电话信道下多说话人识别研究

(申请清华大学工学博士学位论文)

培 养 单 位 ： 计算机科学与技术系

学　　科 ： 计算机科学与技术

研 究 生 ： 邓　　菁

指 导 教 师 ： 吴 文 虎　教　授

副指导教师 ： 郑　　方　研究员

二〇〇六年十月

Studies on Multi-Speaker Recognition over Telephone

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

by

Jing Deng

(Computer Science and Technology)

Dissertation Supervisor: Professor Wen-hu WU

Associate Supervisor: Professor Fang Zheng

October, 2006

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘　要

为了提高电话信道下多说话人识别系统的性能，本论文在背景噪音、多说话人和信道差异三个方面进行了研究。主要工作包括：

1. **提出基于预测差分幅度谱的噪音鲁棒特征的提取算法**。针对电话信道下背景噪音种类繁杂的实际情况，本文提出一种基于帧内相邻频段幅度差的噪音鲁棒的特征提取算法。该算法不需要预先对噪音进行估计，且去噪处理是基于单帧语音内的。对研究常用的四种经典噪音（即 white，babble，f16 和 factory），与常用的非线性谱减法相比，该算法的平均错误率相对下降了 24.1%，并且在不同噪音下，都取得了不错的效果。

2. **提出基于通用背景模型（UBM）的说话人分割聚类算法**。针对电话交谈语音中短语音段较多的情况，本文提出一种基于 UBM 的说话人分割聚类算法。在分割阶段，使用相邻两段语音在 UBM 上的似然比分来找出话者发生切换的时间位置，并通过 BIC 方法对分割结果进行优化。在聚类阶段，将一段语音在说话人模型间的分数差作为该语音段属于某个模型的“概率分”，在对该分数进行 Dnorm 处理后，根据分数的大小，对语音段按话者身份进行归类。为了进一步减少分割阶段产生的错误率，在聚类的基础上，进行了重分割。在 NIST 2002 年 Switchboard 数据库上，分割聚类错误率为 4.5%。相对于当年分割聚类性能最好的系统来说，分割聚类错误率相对下降了 21.1%。

3. **提出基于信道子空间投影的模型补偿算法。**隐藏因子分析（LFA）和干扰属性消除（NAP）是两种效果很好的信道鲁棒算法，但 LFA 的计算非常复杂，而 NAP 不能直接应用于 GMM-UBM 系统中。本文提出一种基于信道子空间投影的模型补偿算法，将 LFA 中模型补偿的思想与 NAP 中子空间投影的思想结合起来，通过子空间投影的方式得到蕴含于语音中的信道信息，并以此对说话人模型进行补偿，提高系统的信道鲁棒性。该算法一方面可简化对语音所蕴含的信道信息的计算，另一方面又可很好地应用于 GMM-UBM 系统中。在 NIST 2006 年说话人识别测试数据库上，该算法与 Tnorm 相结合，等错误率为 9.3%，相对于只用 Tnorm 的基准 GMM-UBM 系统来说，等错误率相对下降了 16.2%。

关键词：多说话人；说话人分割；说话人聚类；噪音鲁棒；信道鲁棒

Abstract

This dissertation focuses on the research on background noise, multi-speaker and channel variance to improve the performance of multi-speaker recognition over telephone. Including:

1. A feature extraction algorithm based on predictive differential amplitude spectrum (PDAS). In order to solve the problem of variant types of noises over telephone, a feature extraction algorithm based on differential amplitudes within one speech frame is proposed. This algorithm does not need advance noise estimation and is performed on a basis of "within one speech frame". Over four types of noises (white, babble, f16 and factory), this algorithm can achieve an average error rate reduction of 24.1% compared with the traditional nonlinear spectral subtraction, and it also performs well under each type of noises.

2. A speaker segmentation and clustering algorithm based on universal background model (UBM). In this paper, a speaker segmentation and clustering algorithm based on UBM is proposed to solve the problem brought by short speech segments in a conversation over telephone. During the segmentation phase, the log likelihood ratio score of two adjecent speech segments between UBM is used as a distance measure to dectect the possible speaker turns in a conversation. After that, BIC is used to refine the segmentation results. During the clustering phase, the differential score of one speech segment between speaker models is viewed as a “probability” to denote how much the speech segment may belong to a speaker model. After Dnorm, these scores are used to cluster speech segments by their numeric value. In order to reduce the error rate introduced by the segmentation phase, re-segmentation is performed by using the results of the clustering phase. On the national institute of standards and technology (NIST) 2002 switchboard corpus, this method achieves an error rate of 4.5%. Compared with the system that achieved the best performance in that year, the relative error rate reduction is 21.1%.

3. A model compensation algorithm based on channel subspace projection.

Latent factor analysis (LFA) and nuisance attribute projection (NAP) are two effecitive channel robustness methods, but the computation of LFA is very complex and NAP can not be used in GMM-UBM system.In this paper, a model compensation algorithm based on channel subspace projection is proposed to analyze the effects of channel variance in supervector space. This method combines the idea of model compensation in latern factor analysis and the idea of subspace projection in nuisance attribute projection together. In order to improve the channel robustness of the system, it uses the channel information, which is estimated from a test utterance by subspace projection, to compensate speaker models whose channel information has already been removed during the training phase. On the one hand, it simplifies the computation of channel information in an utterance; on the other hand, it can be easily used for GMM-UBM systems. On the NIST 2006 single-side one conversation training, single-side one conversation test, this method can achieve an equal error rate of 9.3% when combined with Tnrom. Compared with the conventional GMM-UBM system plus Tnorm, the relative equal error rate reduction is 16.2%.

Keywords: Multi-speaker; Speaker segmentation; Speaker clustering; Noise robustness, Channel robustness

目 录

第 1 章 绪论 1
1.1 说话人识别概述 1
1.1.1 说话人识别系统的研究与发展 3
1.1.2 说话人识别中的特征分析 4
1.1.3 说话人识别中的识别方法 4
1.2 多说话人识别系统的研究现状 5
1.2.1 国内外一些多说话人识别系统与算法简介 6
1.2.2 电话信道下多说话人识别研究的难点 7
1.2.3 研究现状 8
1.3 研究工作概述 11
1.3.1 研究思路 11
1.3.2 工作内容 13
1.4 论文的组织结构 16
第 2 章 噪音鲁棒的说话人识别研究 17
2.1 信号特征级去噪算法 18
2.1.1 谱减法与非线性谱减法 18
2.1.2 差分能量谱 19
2.2 基于预测差分幅度谱的特征提取算法 20
2.2.1 问题的提出 20
2.2.2 基本思想 20
2.2.3 算法描述 21
2.3 实验结果与分析 24
2.3.1 实验设置和数据库 24
2.3.2 实验结果与分析 25
2.3.3 讨论 28
2.4 小结 28

第 3 章　说话人分割聚类研究.....29
3.1 说话人分割聚类研究的现状.....29
3.1.1 基于距离度量的说话人分割聚类算法.....30
3.1.2 基于模型搜索的说话人分割聚类算法.....35
3.1.3 评测指标.....36
3.2 基于 UBM 的说话人分割聚类算法.....38
3.2.1 实验设置和数据库.....39
3.2.2 初始分割.....40
3.2.3 聚类.....46
3.2.4 重分割.....52
3.3 小结.....54
第 4 章　信道鲁棒的说话人识别研究.....56
4.1 常用算法.....57
4.1.1 倒谱均值减.....57
4.1.2 倒谱方差归一.....57
4.1.3 特征弯折.....58
4.1.4 说话人模型合成和特征映射.....59
4.1.5 LFA 和 NAP.....61
4.1.6 Hnorm、Tnorm 和 Znorm.....63
4.1.7 评测标准.....64
4.2 基于信道子空间投影的模型补偿算法.....65
4.2.1 U 矩阵的估计.....67
4.2.2 训练说话人模型.....68
4.2.3 补偿说话人模型.....69
4.2.4 识别测试语音.....70
4.3 实验结果与分析.....72
4.3.1 实验设置和数据库.....72
4.3.2 实验一　单人识别.....74
4.3.3 实验二　双人识别.....77
4.4 小结.....80
第 5 章　总结与展望.....81

5.1 论文工作总结 81
5.2 下一步研究的展望 83
参考文献 85
致谢与声明 94
个人简历、在学期间发表的学术论文与研究成果 95

第1章 绪论

众所周知，语音是人类获取信息的主要来源之一，也是人与外界交流中使用最方便、最有效、最自然的工具。最初人们是通过人耳来辨别语音的话者身份，即“闻声识人”。随着计算机的出现和电子信息技术的发展，出现了用计算机自动识别语音的话者身份的技术，即说话人识别（Speaker Recognition）技术。说话人识别技术有着非常广阔的应用前景：在司法领域，它可以用来协助确认犯罪嫌疑人；在军事领域，它可以用于战场侦听，以辨认敌方指挥员；在银行等处的安全系统中，它可以作为身份核查或安全检查的一种手段；在日常生活中，它可以用作个人身份认证的手段，如声控门、声控命令等等。电话信道下的多说话人识别是上述应用之一，它的目的是要解决电话语音中遇到的背景噪音、多人语音和信道差异等因素所带来的影响，提高多说话人识别系统的性能。论文在前人工作的基础上，对上述三个问题分别进行了研究，并提出了自己的一些见解。

本章的内容安排如下：首先对说话人识别技术的组成与发展做简要介绍；第二节指出电话信道下多说话人识别的重点和难点，并综述其研究现状；最后给出本文工作的研究思路和具体内容。

1.1 说话人识别概述

说话人识别是一项根据语音波形中反映说话人生理和行为特征的语音参数，来识别待测语音话者身份的技术。说话人识别系统，可以简单的定义为：以说话人的语音作为输入，用训练得到的特定人模型来识别待测语音的话者身份。图 1.1 是典型的说话人识别系统的模块示意图，从图中可以看到说话人识别系统的两个组成阶段：训练阶段和识别阶段。在训练阶段，说话人的语音经过特征提取后得到各自的声学特征，然后系统为每个目标说话人建立相应的模型并组成说话人模型库；在识别阶段，用测试语音提取出的声学特征与说话人模型库中的模型进行比较，根据一定的相似性准则来判断测试语音发出者的身份。

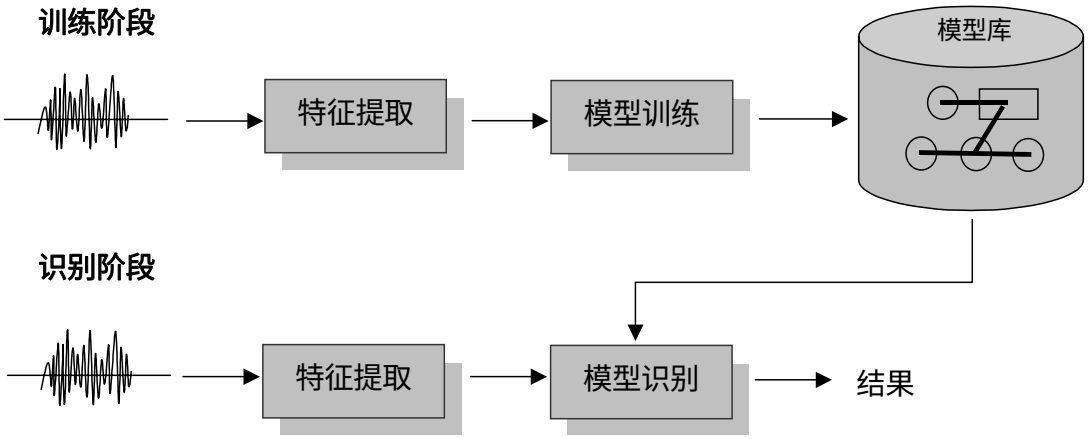

图 1.1 典型说话人识别系统示意图

说话人识别根据应用的范畴可分为两类：(1) 说话人辨认 (Speaker Identification)：把待识别的语句判定为属于 N 个参考说话者中的某一位，是一个多选一的问题；(2) 说话人确认 (Speaker Verification)：确定一段说话人的语句是否与所声明的参考说话人相符，这只有两种选择：或是肯定 (即接收)，或是否定 (即拒绝)。在实际应用中，说话人确认的灵活性和应用性一般要好于说话人辨认，这是因为说话人确认系统允许待测说话人不属于已知的用户集，并且其识别性能不会因为系统用户数量的增加而下降；而说话人辨认系统则要求待测说话人属于已知的用户集，并且其识别性能会随着用户数量的增加而下降。

说话人识别根据识别的内容可以分为两类：(1) 文本相关 (Text-dependent)：在训练时要求用户按照规定的文本发音，精确的建立每个说话人的模型 (例如基于词或音素的模型)；在识别时要求用户也必须按规定的文本发音；(2) 文本无关 (Text-independent)：不规定说话人发音的文本，模型建立相对困难，但用户使用方便，可应用范围较宽。一般来说，文本相关的说话人识别性能要高于文本无关的说话人识别，但是后者使用的灵活性要好于前者。

说话人识别根据测试语音的类别可以分为两类：(1) 开集 (Open-set) 识别：有的待测说话人不属于已知的说话人集合；(2) 闭集 (Close-set) 识别：所有待测说话人都属于已知的说话人集合。显然，说话人开集识别的难度要大于说话人闭集识别。

一般来说，面向电话信道应用的多说话人识别系统的输入语音是任意文本的随意发音，待识别语音中有的属于目标说话人（集内说话人），有的属于假冒者（集外说话人），因此本论文所研究的系统是一个文本无关的开集多说话人识别系统。

1.1.1 说话人识别系统的研究与发展

对说话人识别的研究始于 20 世纪 30 年代，早期的工作主要集中在用人耳进行听辨语音的实验和探讨听音识别的可能性方面。随着电子技术和计算机技术的发展，通过机器自动识别人的语音成为可能。Bell 实验室的 Pruzansky 提出了一种基于模式匹配和概率统计方差分析的说话人识别方法[1]，从而引起信号处理领域许多学者的注意，形成了说话人识别研究的一个热潮。这期间主要工作集中在各种识别参数的提取、选择和实验上。20 世纪 70 年代至今，说话人识别的研究重点转向对各种声学特征参数的线性或非线性处理以及新的模式匹配方法上。如今，说话人识别技术已经逐渐走向实际应用，AT&T 应用说话人识别技术研制出的智慧卡（Smart Card），已经应用于自动提款机上。欧洲电信联盟于 1998 年完成了 CAVE（Caller Verification in Banking and Telecommunication）计划，并于同年启动了 PICASSO（Pioneering Call Authentication for Secure Service Operation）计划，在电信网上完成了说话人识别。其他一些商用系统还包括：ITT 公司的 SpeakerKey、Keyware 公司的 VoiceGuardian、T-NETIX 公司的 SpeakEZ 等。此外，国内许多高科技公司，如中科模识科技公司、中科信利技术有限公司等，也都专门开发了许多说话人识别方面的应用产品。

目前国际上许多著名大学、研究机构以及很多大公司的实验室都在进行说话人识别方面的研究，如麻省理工学院林肯实验室（Lincoln Laboratory）、美国的 ICSI（International Computer Science Institute）、美国的 SRI 公司的语音技术与研究实验室（STAR）、法国的 LIA（Laboratoire Informatique Avigonon）、加拿大的 CRIM（Centre de recherche informatique de Montréal）实验室等。

在国内，许多大学和研究机构也在这一领域开展了大量的研究工作，并在说话人识别方面取得了丰硕的研究成果，如中科院声学所、中

科院自动化研究所、北京大学、中国科技大学、北京邮电大学、北京交通大学、北京理工大学、上海交通大学、哈尔滨工业大学等。

下面本文将主要从特征分析和识别方法两个方面来介绍说话人识别的研究进展。

1.1.2 说话人识别中的特征分析

通常在一段语音信号中包含很多层次的说话人相关信息，这些信息包括低层的声学特征，较高层的韵律、语速和语调等，以及更高层的口音、发音习惯等。如何提取和描述这些信息是进行说话人识别的前提基础。目前常用的特征参数大多数采用的是低层声学特征，例如线性预测倒谱系数（Linear Predictive Cepstrum Coefficient，LPCC）[2]、Mel 频率倒谱系数（Mel-Frequency Cepstrum Coefficient，MFCC）[3,4]和感知线性预测系数（Perceptual Linear Predictive，PLP）[5]等等。虽然分帧处理后的每一帧倒谱参数被认为是独立的，但实际上语音信号的每一帧与其相邻的若干帧之间存在着较大的相关性[6]。常用的处理方法是在静态的倒谱中加入动态信息来强化特征表示，例如加入倒谱的差分和自回归参数等等[7]。此外，一些时域参数和高层信息，例如短时能量、短时能量一阶差分、基音周期、共振峰、习惯用语和基于词或音素的 N 元模型等也常被结合到特征参数表示中来，以提高系统的性能[8~13]。

1.1.3 说话人识别中的识别方法

从识别方法上说，常用的说话人识别方法可分为模板匹配法[14,15]、统计概率模型法、人工神经网络（Artificial Neural Network，ANN）法[16,17]和支持向量机（Support Vector Machine，SVM）[18]等。其中，模板匹配法主要有动态时间归整（Dynamic Time Warping，DTW）法[19]和最小近邻（Nearest Neighbor，NN）法[20]等；统计概率模型法主要有隐马尔可夫模型（Hidden Markov Model，HMM）[21~23]、高斯混合模型（Gaussian Mixture Model，GMM）[24,25]和分段高斯模型（Segmental Gaussian Model）[26]等，其中隐马尔可夫模型和高斯混合模型是说话人识别中最常用的两种概率模型。在文本无关的说话人识别领域，基于高斯混合模型和通用背景模型（Gaussian Mixture Model-Universal

Background Model，GMM-UBM）的说话人识别已经成为主要的识别方法[27,28]；人工神经网络法主要有时延神经网络[29]，决策树神经网络[30]等；基于支持向量机的说话人识别系统中使用的特征也是目前广泛使用的声学特征[31,32]，并且研究人员常常将 SVM 与 GMM 相结合来提高说话人识别系统的性能[33~35]。

在说话人识别中，GMM-UBM 已逐渐成为一种常用的识别方法，其中 UBM 是一个说话人无关、高阶的高斯混合模型。该模型通常由数百人甚至上千人、男女平衡的数小时语音训练得到，用于表示说话人的统计平均发音特性。基于 GMM-UBM 的系统有两个好处：(1) 说话人模型是在 UBM 上根据说话人的训练语音自适应得到的。这样，对于说话人训练语音覆盖到的发音，可以用该说话人自身的语音建模；对于未覆盖到的发音，可以用 UBM 里的发音分布近似，从而减少测试语音与训练语音在声学空间上由于分布不同所带来的影响；(2) UBM 可以被看作是一个“标准参考者”的模型，这样在进行身份确认的时候，可以用测试语音在 UBM 上的得分来作为一种参考阈值。因此，论文中的说话人识别系统是基于 GMM-UBM 的，其中使用的特征是改进的 MFCC（详见第二章）和相应的一阶差分系数。一般来说，常用的自适应算法有最大后验概率（Maximum *A Posteriori*，MAP）算法[36~38]和最大似然线性回归（Maximum Likelihood Linear Regression，MLLR）算法[39~41]。前者需要估计出某一特定环境下的先验模型参数，从而对说话人模型进行相应的补偿；后者假定用一小部分语音数据即可估计出训练环境与测试环境之间在模型参数上的差异，在此基础上，对说话人模型进行修正。在论文使用的实验系统中，说话人模型自适应所使用的算法是 MAP。

1.2 多说话人识别系统的研究现状

随着电子技术和计算机技术的迅速发展，人们可以很容易的获得各种音频文件，例如广播电视节目录音、采访录音、网络音频聊天录音和公安监听录音等等。这些音频文件的数量相当庞大，并随着时间的推移在不断的增多，其中大多数音频文件混有多个说话人的语音、背景音乐和环境噪音等等。这样，就要求研究一种能够自动的从一堆音频文件中

找到所需特定说话人语音的技术，这一技术被称为多说话人识别，或者说话人检测或跟踪（Speaker Detection or Tracking）。多说话人识别主要要解决两个问题，即谁在说话，和在什么时候说话。“谁在说话”这一问题是由说话人确认技术来完成，而“在什么时候说话”则由分割（Segmentation）和聚类（Clustering）技术来解决，在有的文献中，将分割和聚类合并到一起，统称为说话人分割。

1.2.1 国内外一些多说话人识别系统与算法简介

经过国内外研究机构一段时间的潜心研究，目前已经出现了不少具有实用价值的多说话人识别算法和系统，以下是其中一些系统的简介。

一、法国的 ELISA 系统将 CLIPS 系统和 LIA 系统进行融合，用于多说话人检测。其中 CLIPS 系统是一个基于贝叶斯信息准则（Bayesian Information Criterion，BIC）[42]的多说话人检测系统，LIA 系统是一个基于 HMM 的模型搜索的多说话人检测系统。ELISA 系统参加了多次美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）组织的说话人分割聚类评测，在 NIST 2002 评测中，ELISA 系统在会议交谈语音库和电话对话语音库上取得了最优性能[43]；在 NIST 2003 评测中，ELISA 系统获得了最优系统性能[44]；在 NIST 2004 评测中，ELISA 系统取得了最优说话人分割性能[45]。

二、美国 AT&T 实验室提出了一种基于 GMM 的说话人检测算法[46]。该算法首先根据训练语音得到目标说话人模型与背景说话人模型，并根据语音段在这些模型上的似然分数差，来进行目标说话人检测。在 HUB4 新闻数据库上，对于单一目标说话人检测，在语音质量很干净的情况下，漏检率大约是 7%；在语音质量不干净的情况下，漏检率大约是 27%。

三、微软亚洲研究院提出了一种基于 UBM 的说话人实时分割算法[47]。该算法分为两步：预分割和优化。在预分割阶段，根据每帧语音在 UBM 上的得分大小将该帧语音划分为可靠说话人语音帧、可疑说话人语音帧和非说话人语音帧；在优化阶段，使用递增说话人自适应（Incremental Speaker Adaptation，ISA）算法从可靠说话人语音帧上得到精确的说话人模型，并根据得到的模型对初始分割的结果做进一步判

决。在 HUB4 英语新闻广播数据库上，误警率为 19.23%，漏检率为 13.65%。

四、法国 P. Delacourt 等人提出了一种叫做 DISTBIC 的说话人分割算法[48]。该算法分为两步：预分割和优化。在预分割阶段，采用一般化似然比（Generalized Likelihood Ratio，GLR）[26,49]和 KL 距离（Kullback-Leibler Distance）[50]进行初始分割；在优化阶段，使用 BIC 来判断初始分割中相邻两个语音段是否属于同一个说话人，如果是则合并，否则保持不变。该算法在新闻语料和电话语料上都取得了不错的分割结果。

五、北京大学信息科学技术学院智能科学系的视觉与听觉信息处理国家重点实验室提出了一种基于集外模型集上分数向量的说话人分割算法[51]。该算法包括预分割、集外模型算分和基于模型分数向量的聚类三部分，在 NIST 2003 双说话人识别数据库上取得了较好的分割效果。

六、中国科学院自动化研究所高技术创新中心提出了一种基于熵的音频跳变点检测方法[52]，切分后的语音片断通过说话人聚类来重新定位语音中说话人的变化点。新的语音片断，经过维纳滤波和 Pitch 端点检测，用于最终的说话人检测系统。在广播电视音频流上的错误率相对于 BIC 方法降低了 7.9%。

七、中国科学院声学研究所与中科信利实验室构建了一个完整的广播新闻语料识别系统（ThinkIT-BNR）[53]。该系统包括：音频匹配、音频自动分段、音频分类、说话人聚类、识别后处理和多阶段识别策略等多个模块，对新闻联播节目的误识率为 10.14%。

1.2.2 电话信道下多说话人识别研究的难点

对于实验室环境下录制的干净语音，说话人识别系统一般都能达到较高的识别率。但是由于电话信道下的多说话人识别系统的应用背景十分复杂，往往（1）输入的语音通常会伴随着一定的环境噪音，并且不同的说话场所噪音的类型也不尽相同，给说话人识别增添了难度；(2) 由于电话信道畸变、移动电话和固定电话传输信道特点不同、以及不同的采音设备（如不同的手机类型、不同的座机型号等），都会对语音信

号产生一定的影响，这将直接影响到最终的识别性能；(3) 由于应用场景是电话信道，不可避免的会遇到多人同时发音或者说话人发生切换的情况，使得输入语音包含有多个说话人信息，给说话人识别系统带来很大的困难；(4) 面向电话应用的说话人识别系统要求能够对输入语音进行身份判决的问题，即哪些语音属于目标说话人，哪些语音属于假冒者。在上述四个问题中，第一点和第二点会较大地降低说话人识别系统的性能；第三点是电话信道下多说话人识别系统要解决的问题；第四点是面向实际应用的说话人识别系统必须解决的问题。

综上所述，背景噪音、信道差异、多人语音和说话人拒识是目前电话信道下多说话人识别系统中的难点，其中对于多人语音的处理一般称为说话人分割聚类。

1.2.3 研究现状

下面将简单介绍一下噪音鲁棒、信道差异、说话人分割聚类和说话人拒识四个方面的研究现状。

1.2.3.1 噪音鲁棒

我们知道，语音质量的好坏，将会直接影响说话人识别系统的性能。而在实际应用中，由于说话者可能处于各种各样的环境，这将会使得录制的语音受到不同类型噪音的影响，从而降低系统的识别性能。因此噪音鲁棒性问题一直是说话人识别研究中的热点和难点问题之一。对噪音鲁棒的研究通常可以分为两大类：(1) 信号特征级去噪：该类算法主要是从信号处理的角度出发，或者去除噪音的影响，或者提高特征对噪音的抗干扰性。它包含了语音信号检测、噪音消除、信噪分离以及语音信号增强等众多技术。常用的算法有谱减（Spectral Subtraction，SS）法[54~56]、非线性谱减（Non-linear Spectral Subtraction，NSS）法[57,58]、RASTA 滤波法[59,60]、E-RASTA 法[61]、正则相关分析的谱变换补偿（Canonical Correlation Based on Compensation，CCBC）法[62]、主分量分析（Principal Component Analysis，PCA）法[63]、异方差线性可区分性分析（Heteroscedastic Linear Discriminant Analysis，HLDA）法[64]等；(2) 模型级去噪：该类方法主要是在声学模型级上研究噪音问题，通

过模型补偿（Model Compensation）技术[65,66]，减少测试集和训练集的不匹配，从而提高系统对含噪语音的识别性能。

1.2.3.2 信道差异

在实际应用中，由于说话者使用的设备不同（如不同型号的手机、座机等）、或者传输信道（如 GSM、CDMA、小灵通等）的不同，导致录制的语音受到不同程度的影响，这些影响统称为信道差异。信道差异的存在会使得测试语音和训练语音之间存在一定的不匹配，从而降低说话人识别系统的性能。因此，对信道差异这一问题解决的好坏将直接影响到说话人识别系统能否投入实际应用。目前对信道差异的研究可以分为三个方面：(1) 特征域：该类算法从信号处理的角度出发，或者消除信道对声学特征的影响，或者提高特征对信道的鲁棒性。常用的算法有倒谱均值减（Cepstral Mean Subtraction，CMS）[67]、RASTA 滤波、特征弯折（Feature Warping）[68]、特征映射（Feature Mapping）[69]等；(2) 模型域：该类算法主要是在模型空间对信道进行补偿或消除，以便消除训练环境与测试环境的不匹配。常用的算法有说话人模型合成（Speaker Model Synthesis，SMS）[70]、HNSSM（Handset Normalization in Synthesized Speaker Model）[71]、隐藏因子分析（Latent Factor Analysis，LFA）[72,73]、干扰属性消除（Nuisance Attribute Projection，NAP）[74,75]等；(3) 分数域：该类算法主要是预先估计出假冒者语音得分的分布（通常为单高斯分布），然后用该分布对测试语音的得分进行归一化处理，来减小信道差异对语音得分的影响。常用的算法有 Hnorm[76]、HTnorm、Cnorm[69]、基于语音编码差异的似然比得分补偿等[77]。目前多数研究者使用的信道鲁棒算法往往是上述算法中的某个或者某几个的融合。

1.2.3.3 说话人分割聚类

在基于电话信道的实际应用中，话者不可避免的会发生切换，或者由于设备的问题，不能将对话双方的语音自动分离，从而使得输入语音含有多个说话人，从而影响到说话人识别系统的性能。由于目前大多数说话人识别算法是针对单人语音的，无法直接处理多人语音，因此这就

要求研究一种能够将多人语音转变为多段单个语音的技术，该技术称为说话人分割聚类。对说话人分割聚类的研究主要可以分为以下两类：(1) 基于距离的：该类算法主要是利用一些距离度量准则，判断两段语音是否是同一说话人所说。常用的算法有 GLR、KL 距离、BIC、DISTBIC 等；(2) 基于搜索的：该类算法主要是根据先验的目标说话人模型或从多人语音流中估计得到的说话人模型，采用搜索的方式来判断目标说话人在什么时候发音。常用的算法有基于 GMM 的搜索算法、基于 HMM 的搜索算法等。近些年来，一些多说话人识别系统将二者有效地结合起来，取得了较好的效果，如法国的 ELISA 系统等。其他较常用的准则还有过零率比（Zero-Crossing Rate Ratio，ZCRR）、短时能量比（Short-Time Energy Ratio，STER）和频谱流（Spectrum Flux，SF）[78] 等。虽然在设备条件允许下，利用麦克风阵列可以更好的对含有多个说话人的语音流进行处理；但对于大多数多说话人识别的应用来说，一方面系统事先无法控制和预测用户使用的设备类型和应用场景，另一方面麦克风阵列的成本较高，因此在大多数多说话人识别应用中，使用最多的还是基于距离度量或模型搜索的算法。

1.2.3.4 说话人拒识

面向实际应用的说话人识别系统，要求能够对输入语音的话者身份进行判别，如果是目标说话人则接受，如果是假冒者则拒绝。一般来说，文本相关的说话人确认系统的性能要好于文本无关的系统，但前者的应用灵活性要比后者差很多。在银行、超市等远程电话服务领域，由于用户对说话人识别系统的拒识性能要求很高，而且也比较乐于或能够较好的配合系统要求，因此这些领域的说话人识别系统常采用基于文本相关或者文本提示的方式，将语音识别技术与说话人识别技术相结合，来提高系统的拒识性能[79~81]。而在公安监听等领域，由于话者不可能按照固定的文本进行发音，因此这些领域的说话人识别系统常采用文本无关的方式来确认话者的身份。我们知道，基于 GMM-UBM 的系统本身就可以作为一个简单实用的文本无关的说话人确认系统，在信道已知的情况下，可以取得较好的拒识性能。为了提高系统的拒识性能，研究人员

对语音得分进行了归一化处理[82]，将目标说话人的得分与假冒者的得分区分开来。常用的分数归一化方法有 Znorm[83]、Tnorm[84]和 Dnorm[85]等。为了进一步提高说话人确认系统的拒识性能，近年来许多研究人员加大了对语音中高层信息的研究力度，以求能够更好地和更多地从语音中提取出跟特定说话人相关的信息，如基频和能量的分布[86]、韵律统计（Prosodic Statistics）[87]、音素 *N* 元模型（Phone *N*-gram）[88]、发音模型（Pronunciation Modeling）[89]、词 *N* 元模型（Word *N*-gram）[90]等。研究表明，将高层信息与低层信息有效的结合，在一定条件下能够提高说话人确认系统的拒识性能。

1.3 研究工作概述

1.3.1 研究思路

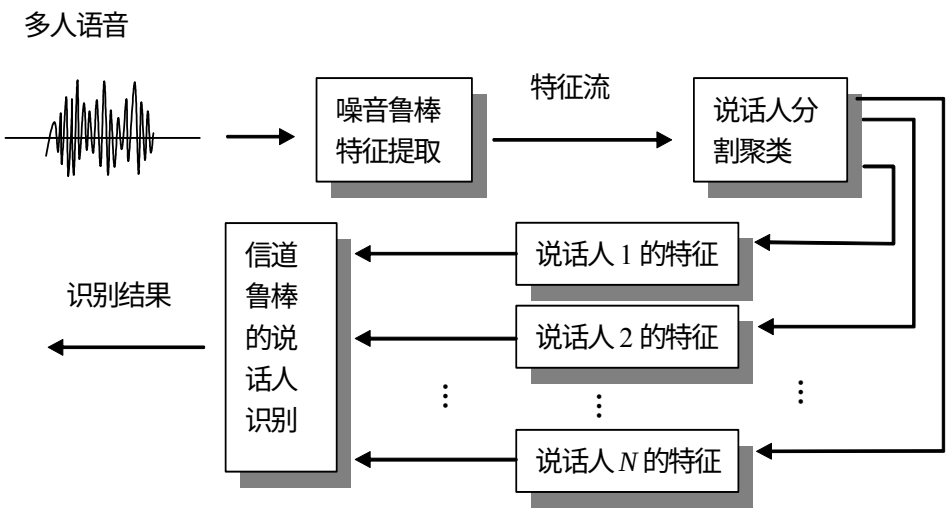

图 1.2 多说话人识别研究思路示意图

面向电话信道应用的多说话人识别系统，可以采用分割聚类的方式来将含有多个说话人的语音流分解为多段只含有单个说话人的语音流，从而将多人识别任务转变为多个单人识别任务。然而在这类应用中，常会受到背景噪音和信道差异的影响，它们会较大的降低说话人分割聚类算法的性能。由于电话交谈语音中的每个说话人所处的信道类型一般是不同的，且不易在分割聚类前对交谈语音中的信道差异进行处理；而对

于电话交谈语音中含有的噪音，如果不在分割聚类前进行一定的去噪处理，则会直接影响分割聚类算法的性能。综上所述，本文采用的研究思路（如图 1.2 所示）为：在进行分割聚类前，先对原始语音流中含有的噪音进行处理，来得到噪音鲁棒的特征流；在噪音鲁棒的特征流基础上，首先找出说话人发生转换的时间位置（分割），然后按照话者的身份对分割后的语音段进行归类（聚类）；在得到多段只含有单个说话人的特征流后，对每段特征流进行信道差异的消除或补偿，来降低训练语音与测试语音之间的信道不匹配。

从图 1.2 中可以看到，面向电话信道的多说话人识别被分解为三个前后关联的研究问题，对这三个问题的研究思路如下：

（1）噪音鲁棒特征的提取：由于电话交谈语音中含有的噪音类型是未知的，因此难于预先得到噪音谱的准确估计，使得谱减法的应用有一定的困难。而 Chen[91]等人提出的差分能量谱算法不需要预先对噪音谱进行估计，通过在原始语音谱上计算差分能量谱来提高系统的噪音鲁棒性。但是差分能量谱破坏了原始语音流在频谱中的峰谷信息，会在一定程度上影响说话人识别系统的性能。因此，本文在差分能量谱的基础上进行了改进，以便尽可能的保留语音谱中的峰谷信息，来提高说话人识别系统的识别率；

（2）说话人分割聚类：由于电话交谈语音中存在较频繁的话者切换现象，即交谈语音中含有较多的短语音段，这些短语音段会给分割聚类算法带来一定的困难。这是因为，说话人分割算法大多是基于语音窗分析的，要求说话人一次发音的时间不能太短，否则在单个语音窗内就会出现两次以上的话者切换，影响算法的性能。考虑到在较短的时间段内，同一个说话人的两段语音在一个高精度模型上的似然分差异不大，并且由于模型精度足够高，能够较好的对短语音段进行身份识别，因此可以用语音段在一个高精度模型上的似然比分来作为一种分割距离度量准则，来解决短语音段较多的问题。由于 UBM 代表了大多数说话人的发音特性，是一个高精度的模型，并且语音段在 UBM 上的似然分差异也能够在一定程度上反映出它们之间在声学分布上的差异，因此，在分割阶段，本文将语音段在 UBM 上的似然比分作为一种距离度量准则，用以找出交谈语音中可能的说话人转换点；在聚类阶段，由于电话交谈

语音一般只含有两个说话人，因此本文对所研究的问题做了简化，假定多人语音中只含有两个说话人。为了对语音中含有的两个说话人建立正确的说话人模型，并在此基础上对分割后的语音段按照话者的身份进行归类，本文提出了一种基于模型间分数差的聚类方法。该算法将一段语音在两个模型上的分数差作为该语音段属于其中某个模型的“概率分”，并根据“概率分”的大小，选择得分较大的语音段用于训练或更新特定说话人的模型，以保证训练用语音段的话者身份尽可能一致。由于聚类没有改变分割的结果，使得在分割时产生的漏检错误不能得到一定的消除，因此，本文在聚类后进行了重分割处理，利用聚类时得到的说话人模型，对多人语音按窗进行重新分割。这样，通过说话人分割聚类，原始的多人语音就改变为多段只含有单个说话人的语音流；

（3）信道鲁棒的说话人识别：LFA 和 NAP 是近年来在说话人识别领域提出的非常有效的信道鲁棒算法，它们都是在超向量空间上分析信道差异给语音带来的影响，其中 LFA 是基于模型补偿的算法，应用于 GMM-UBM 系统中；而 NAP 则是基于消除模型中干扰说话人识别的信息的算法，应用于 GMM-SVM 系统中。但是，LFA 算法的计算复杂度很大，应用起来比较困难；而 NAP 算法则不能直接应用于 GMM-UBM 系统中。本文在这两种算法的基础上，将 LFA 中模型补偿的思想与 NAP 中子空间投影的思想结合起来，其基本思想是用子空间投影的方式来得到语音中含有的信道信息，并将得到的信道信息补偿到说话人模型上，以便降低测试语音和训练语音间由于信道差异所带来的不匹配。这样，本文提出的算法一方面简化了对语音中信道信息计算的复杂度，另一方面又能够较容易的应用于 GMM-UBM 系统中。

在对上述三个问题的研究基础上，我们可以构建一个完整的多人识别系统。虽然这只是一个初步的系统，里面还存在很多需要进一步研究的问题，但仍然可以在一定条件下用于实际应用中。

1.3.2 工作内容

论文的研究工作涉及到电话信道下多说话人识别系统面临的背景噪音、多人语音和信道差异三个方面的问题，如图 1.3 所示。

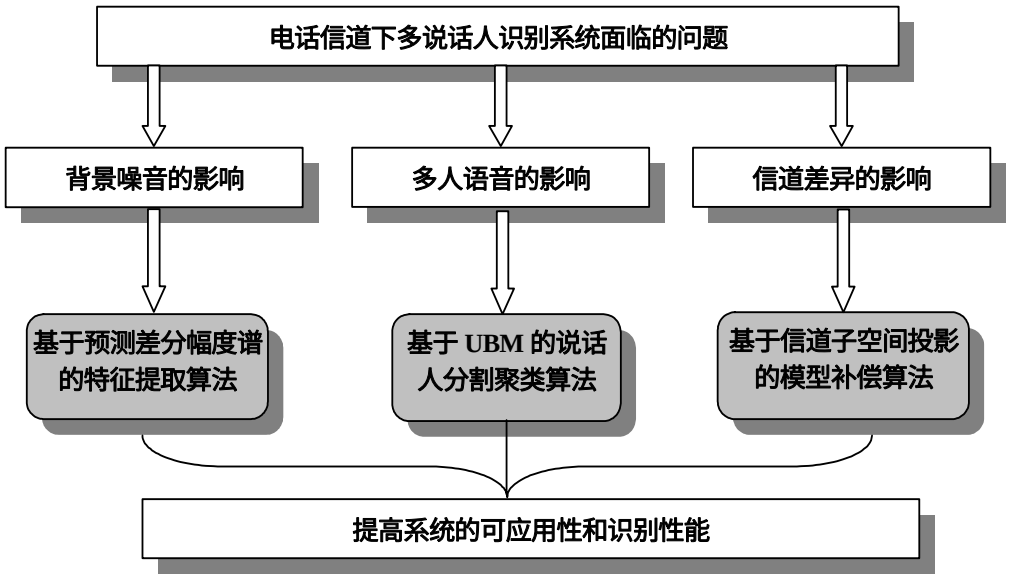

图 1.3 作者的工作内容

具体地说，作者的工作内容包括以下几个方面：

1) 针对电话语音中噪音类型的多样性的问题，提出一种基于预测差分幅度谱（Predictive Differential Amplitude Spectrum，PDAS）的噪音鲁棒的特征提取算法。由于语音频谱中的峰谷信息包含有很强的说话人特性信息，而背景噪音会导致测试语音频谱中的峰谷信息与训练语音不匹配，从而降低系统的识别性能。为了减少由背景噪音带来的这种不匹配，本文在差分能量谱的基础上提出了一种基于 PDAS 的噪音鲁棒特征提取算法，利用一帧语音内相邻频率上的幅度差来去除噪音的影响。由于噪音的存在，会使得语音频谱中原来清晰的峰谷信息变得模糊，因此，本文首先通过使用一个正弦滤波器来预测（或估计）当前语音帧频域内的峰谷信息，然后采用不同的加权差分函数来去除背景噪音的影响，最后通过累积运算来恢复原始干净语音的频谱。在混有 White、F16、Babble 和 Factory 四种噪音的不同信噪比的语音中，与非线性谱减法和差分能量谱法进行了比较分析。

2) 针对电话信道下多人语音中的短语音段较多的情况，在噪音鲁棒特征下，提出一种基于 UBM 的说话人分割聚类算法。该算法包括三个阶段：初始分割、聚类和重分割。说明如下：(1) 在初始分割阶段，利用 UBM 能够较好的描述说话人发音共性的特点，根据相邻两段语音

在 UBM 上的似然比分来找出多人语音中话者发生切换的时间位置。为了减少分割算法产生的错误分割点,使用 BIC 方法对分割后相邻两段语音进行合并判断，在一定程度上降低了分割算法的误警率，但漏检率有稍许上升；(2) 在聚类阶段，提出了一种基于模型间分数差的聚类方法。该算法将一段语音在两个说话人模型上的分数差作为该语音属于某个模型的“概率分”，并按照“概率分”的大小，选择数值较大的语音段来训练或更新特定说话人的模型。该阶段又可分为两个小阶段：初始聚类和迭代归类。在初始聚类阶段，由于分割后的语音段中含有较多的短语音段，因此使用小混合的 UBM 来估计说话人模型，这样模型的精度能够得到保证；在迭代归类阶段，由于可用于训练说话人模型的语音足够长，可以使用较大混合的 UBM 来得到精度更高的说话人模型，以保证聚类结果的准确率；为了降低说话人自身差异给语音得分带来的影响，对语音得分做了 Dnorm 处理，进一步降低了算法的分割聚类错误率；(3) 在重分割阶段，由于分割算法引入的漏检错误，不能在聚类时加以消除,因此用聚类时生成的说话人模型对原始语音流进行了重新分割。论文提出的算法在 NIST 2002 Switchboard 数据库上（电话交谈语音库），取得了较好的分割聚类效果，并跟 NIST 2002 年说话人分割聚类评测中，该数据库上分割聚类性能最好的系统进行了比较。通过说话人分割聚类后，原始的多人语音就改变为多段单人语音，也就是说，原先的多人识别任务可以通过多个单人识别来完成。

3) 针对电话信道应用中信道类型多样性的问题，提出了一种基于信道子空间投影（Channel Subspace Projection，CSP）的模型补偿算法。该算法将测试语音中的信道信息(即测试语音在信道子空间上的投影)，补偿到训练语音生成的说话人模型(已去除了训练语音在信道子空间上的投影）上，以减轻信道不匹配所带来的影响。具体步骤如下：(1) 首先用 PCA 方法从大量的集外说话人数据上得到信道子空间的一组正交基，并构建投影矩阵；(2) 在训练时，首先用训练语音从 UBM 上得到说话人的模型，然后去除该说话人模型在信道子空间上的投影。同样，对 UBM 也进行相同的处理，记做 UBM'；(3) 在测试时，首先用测试语音从 UBM 上得到说话人的模型，然后计算该说话人模型在子空间上的投影，并用该投影对 UBM'和说话人模型（在训练阶段得到的说话人

模型）进行补偿；最后用补偿后的 UBM'和说话人模型来对测试语音进行话者身份的识别。在 NIST 2006 年单说话人识别数据库上，对该算法进行了测试，并与基准系统进行了比较分析。同时，构建了一个完整的多人识别系统，将论文所提出的三个算法整合到一起，在 NIST 2006 年的多人识别数据库上，测试了系统的性能。

1.4 论文的组织结构

本文的内容共五章，具体安排如下：

第一章是绪论部分，首先概述了国内外说话人识别技术的研究发展历史和现状；接着综述了国内外多说话人识别的研究现状和难点问题，并在此基础上阐述了研究的思路和工作内容；最后是论文的内容安排介绍。

第二章先介绍噪音鲁棒性方面的相关研究，并针对电话语音中背景噪音的特点，提出基于预测差分幅度谱的特征提取算法，用来提高系统的噪音鲁棒性；然后通过实验分析比较了不同噪音鲁棒算法的性能。

第三章先介绍国内外说话人分割聚类方面的相关研究，针对电话交谈语音的特点和存在的问题，提出一种基于 UBM 的说话人分割聚类算法。该算法包括三个步骤：初始分割、聚类和重分割。通过分割聚类后，将多人语音变为多段单人语音，以便于后面的识别模块进行处理。最后结合实验分析了该算法的有效性。

第四章先介绍国内外使用的信道鲁棒说话人识别算法，接着提出了基于信道子空间投影的模型补偿算法，该方法通过估计测试语音在信道子空间的投影，来对 UBM 和说话人模型进行补偿。最后通过实验比较分析了各种算法的性能，并在多人测试数据库上测试了本文使用的多说话人识别系统。

第五章是总结和展望部分，给出了论文所做工作和成果的总结，并指出了研究中存在的不足之处和对相关领域研究的展望。

第2章 噪音鲁棒的说话人识别研究

面向电话信道应用的说话人识别系统，不可避免地会遇到各种各样的环境噪音，而这些噪音将会使得测试语音与训练语音之间出现不匹配，从而影响系统的识别性能。因此，对背景噪音的处理就成为说话人识别投入实际应用所必须解决的问题之一。一般来说，噪音所带来的失配可以映射到信号、特征和模型三个空间，这里我们将信号与特征空间的噪音鲁棒算法称为信号特征级噪音鲁棒算法，而将模型空间的噪音鲁棒算法称为模型级噪音鲁棒算法。下面对这两类噪音鲁棒算法做简单介绍：

1．信号特征级噪音鲁棒算法

信号特征级噪音鲁棒算法目的是为了消除含噪语音信号中的噪音成分，减少测试语音和训练语音之间因噪音带来的不匹配。主要有以下一些方法：谱减法、子带谱减法（Sub-band SS）、最小均方误差估计（Minimum Mean Square Error，MMSE）[92]、谐波分析、子空间分解[93]、RASTA、特征加权[94]、倒谱均值减法、倒谱归一化法（Cepstrum Normalization，CN）[95]、维纳滤波（Weiner Filter）[96]和基于人耳听觉特性的稳健特征[97]等。另外，一些研究人员将多种特征进行融合，在噪音环境下，取得了较好的识别性能[98]。

2．模型级噪音鲁棒算法

模型级噪音鲁棒算法着眼于调整统计模型的参数，使得模型和含噪语音相匹配。主要的方法有：并行模型合并（Parrallel Model Combination，PMC）[65]、加权投影法（Weighted Projection Measure，WPM）[99]等。这类算法利用了语音和噪音的统计知识，对语音模型进行补偿，来提高系统的识别性能。

上述的一些算法虽然是应用于语音识别中的，但仍然可以被说话人识别所借鉴，来提高说话人识别系统的噪音鲁棒性。

本章主要研究的是声学特征一级的噪音鲁棒特征提取算法，即属于信号特征级噪音鲁棒算法。本章内容安排如下：第一节简单介绍实验里使用的几种信号特征级去噪算法。第二节介绍本文提出的基于预测差分

幅度谱的噪音鲁棒特征提取算法；第三节给出几种算法在四种噪音、不同信噪比环境下的实验结果和比较分析；最后给出本章的小结。

2.1 信号特征级去噪算法

2.1.1 谱减法与非线性谱减法

谱减法假定语音和噪音相互独立，首先估计出噪音能量谱，然后从含噪语音的能量谱中减去噪音的能量谱来得到干净语音的能量谱。假设含噪语音 $y(n)$由干净语音 $s(n)$和噪音 $b(n)$组成，如下式所示：

$$y(n)=s(n)+b(n) \tag{2-1}$$

在频域里，由于事先假设语音跟噪音无关，因此含噪语音能量谱可以表示为：

$$Y(k)=S(k)+B(k),\quad 1\le k\le N \tag{2-2}$$

其中，N 是 FFT 的窗长。

Boll[54]在 1979 年提出的谱减法中，将无语音段的能量谱平均值作为噪音能量谱的估计，记为 $\hat{B}(k)$。干净语音能量谱可以通过下式得到：

$$\hat{S}(k)=Y(k)-\hat{B}(k) \tag{2-3}$$

式（2-3）会使得最终的能量谱出现负值，一般将其置为零值，然而这样的处理会产生频谱尖刺，导致所谓的“音乐噪音（Musical Noise）”。为了解决这一问题，Berouti[55]提出了改进的算法，其出发点是使得谱减后出现负值的机会最小。用公式描述如下：

$$\hat{S}(k)=\begin{cases} Y(k)-\alpha\hat{B}(k), & \left|Y(k)-\alpha\hat{B}(k)\right|>\beta\hat{B}(k) \\ \beta\hat{B}(k), & \text{其他} \end{cases} \tag{2-4}$$

其中 α 是“过估计”(Over-Estimation)因子，β 是“谱底”(Spectral Floor)因子。一般来说，α>1，0<β<<1。为了达到更好的效果，α 可以定义为信噪比函数，信噪比高的部分需要较小的补偿，信噪比低的部分需要较

大的补偿。

上述方法中的噪音能量谱在估计出来后，就保持不变，这并不符合实际应用中的噪音谱的分布情况。为了解决这一问题，Poruba[58]对噪音估计进行了扩展，采用一种非线性的处理方式，用公式描述如下：

$$\hat{B}_i(k)=\lambda\hat{B}_{i-1}(k)+(1-\lambda)B_i(k) \qquad (2\text{-}5)$$

其中 $\hat{B}_i(k)$表示第 i 帧的噪音谱估计，$B_i(k)$表示第 i 帧的噪音能量谱，λ一般取决于当前帧的信噪比。本文中使用的谱减法是基于式（2-4）和式（2-5）的非线性谱减法。

由于噪音叠加的不可恢复性，谱减法会破坏原始语音的频谱。特别是当噪声较大时，谱减法往往会在抹去噪音的同时，也将语音抹去了许多，导致原始语音中有用信息的丢失，从而影响系统的识别性能。另外谱减法需要做预先的噪音估计，对于电话语音中含有多种类型噪音的情况，解决起来就比较困难。

2.1.2 差分能量谱

Chen 等人提出了一种基于差分能量谱（Differential Power Spectrum，DPS）的噪音鲁棒特征提取算法[91]。基于 DPS 的特征（DPS-based Cepstral Coefficients，DPSCC）提取过程可以参看图 2.1：

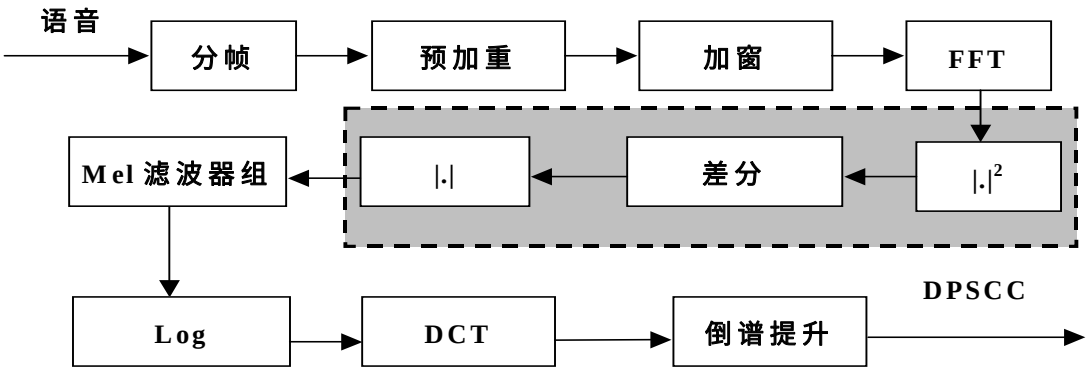

图 2.1 DPSCC 提取示意图

图 2.1 中灰色区域是与传统 MFCC 提取过程不同的地方，在经过 FFT 之后，使用如下的差分函数

$$D(k)=Y(k)-Y(k+1) \tag{2-6}$$

来求得差分能量谱。由于差分运算会产生正负值，所以在差分运算后，使用取绝对值运算来将负值变为正值。然后按照传统的 MFCC 提取算法计算噪音鲁棒特征 DPSCC。

基于差分能量谱的算法的好处是不需要对噪音谱进行预先估计，并且通过实验证明了差分能量谱对噪音的抗噪性要好于原始能量谱。但是由于实际噪音对语音不同频率的分量影响不同，通过简单的差分运算并不能很好的消除噪音的影响，尤其是在较低的信噪比下。

2.2 基于预测差分幅度谱的特征提取算法

2.2.1 问题的提出

电话语音由于应用场景的多变性常会受到各种噪音的影响，如地铁或机场的喧闹声、背景音乐音和话筒质量不好带来的噪音等。如此多样的噪音类型使得噪音能量谱的估计变得非常困难，导致谱减法的性能下降；基于差分能量谱的噪音鲁棒算法不需要进行噪音谱估计，是一种可行的噪音鲁棒算法，但是由于差分能量谱模糊了原始语音在频域上的峰谷信息，使得语音中包含的说话人的特性信息也变得模糊；并且差分能量谱使用了简单的差分函数，使得含噪语音上得到的差分能量谱与干净语音得到的差分能量谱有较大的差别。为了尽量保留原始语音在频域中的峰谷信息，在差分能量谱的基础上，作者提出了一种基于预测差分幅度谱的特征提取算法。

2.2.2 基本思想

由于语音频谱中的峰谷信息包含有很强的说话人特性信息[100]，而背景噪音对原始干净语音谱中波谷的提升要比波峰明显，也就是说噪音对波谷的影响要大于波峰，从而使得原先清晰的峰谷包络变得模糊，导致一些说话人特性信息的丢失。如果能够较好地从含噪语音中还原出原始干净语音谱的峰谷信息，就能提高说话人识别系统对噪音的鲁棒性。针对这一问题，作者提出了一种利用正弦滤波器来预测（或估计）当前

频率位置是处于波峰还是波谷的思路。在得到峰谷信息之后，采用加权的差分函数按照从低频到高频和从高频到低频两个方向来求取右向和左向差分幅度谱，最后采用累积运算（Integral Operation）从差分幅度谱中还原干净语音谱。

2.2.3 算法描述

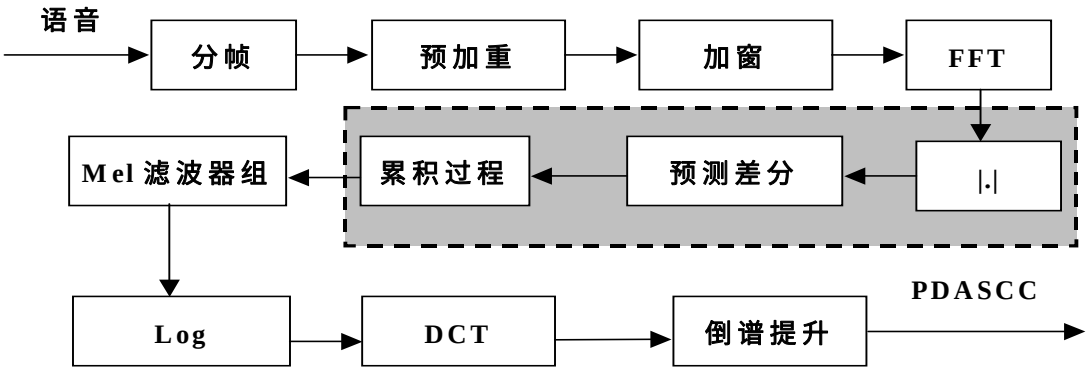

图 2.2 PDASCC 提取示意图

图 2.2 给出了基于预测差分幅度谱的特征（PDAS-based Cepstral Coefficients，PDASCC）的提取过程，其中灰色部分是与传统 MFCC 提取过程不同的地方，包括预测差分和累积过程两步，下面详细介绍这两步。

（1）预测差分：

一帧语音经过 FFT 后，我们可以得到该帧语音的幅度谱。该幅度谱是由许多峰和谷组成的，这些峰谷信息含有相当强的说话人特性信息。然而噪音的影响会使得原先清晰的峰谷变得不明显，从而造成测试语音与训练语音的不匹配。因此，在计算差分幅度谱的时候，需要特别注意频谱中峰谷的位置，以便较好的计算出幅度差分。在频谱中需要注意四种峰谷位置，即峰的左右两侧、谷的左右两侧，如图 2.3 所示。

图 2.3 是干净语音频谱中上述四种峰谷位置的示意图，其中（a）和（b）对应于波峰的左右两侧，（c）和（d）对应于波谷的左右两侧，每个子图有四个点：k、k+1、k+i 和 k+i+1（k-1、k、k+i-1 和 k+i），根据这四个点处幅度值的大小关系，就可以知道 k 和 k+1（k-1 和 k）是处于一个峰点，还是一个谷点。这里 i 是一个比较小的值，实验中它介于

1 到 6 之间。然而背景噪音的影响，会破坏干净语音频谱中 k、k+1、k+i 和 k+i+1（k-1、k、k+i-1 和 k+i）上幅度值之间的大小关系，为了更好地估计含噪语音谱中的峰谷位置，本文使用了一个正弦滤波器来估计频率点 k 和 k+1 右侧附近可能的幅度值，以此来预测 k 和 k+1 所处的位置。

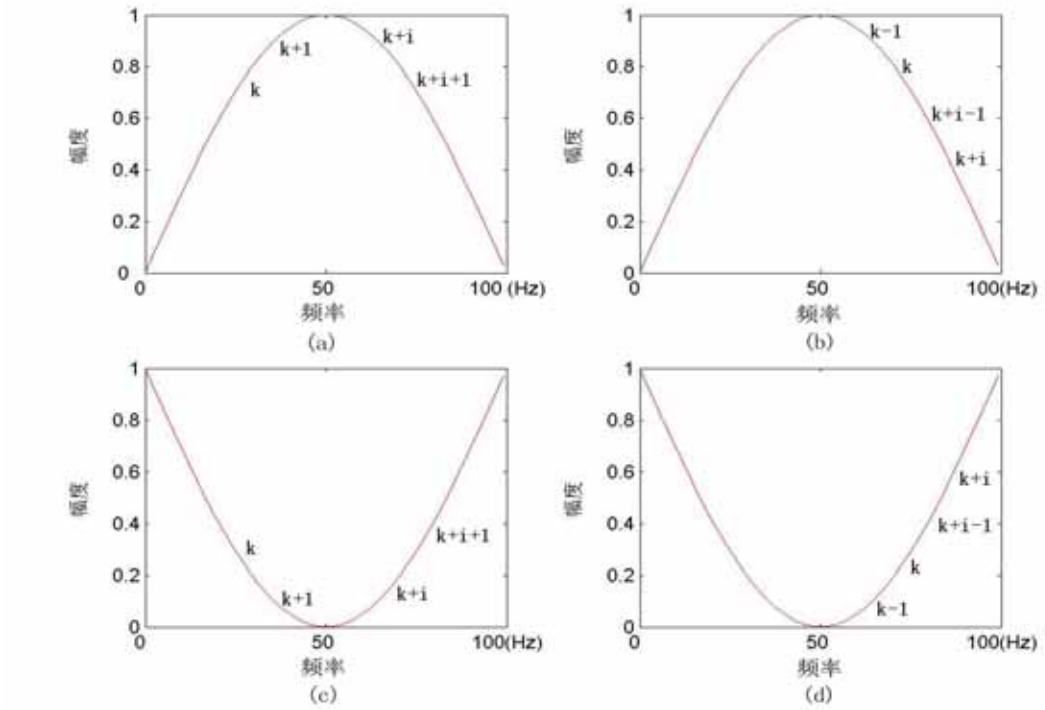

图 2.3 干净语音频谱中的四个峰谷位置

该正弦滤波器的定义如下：

$$h(i)=\sin\left(\frac{\pi}{2}\cdot\frac{i}{W}\right),\quad 0\le i\le W \tag{2-7}$$

其中 W 是滤波器的窗长。那么 k 和 k+1 附近可能的幅度值可以按照下面的公式估计得到：

$$A'(k)=\arg\max_{i}\left[Y(k+i)\cdot h(i)\right] \tag{2-8}$$

这里，$A'(k)$是对 $Y(k)$右侧附近可能的幅度值的估计，即在 k+i 上可能的幅度值。

在得到每个频率点后可能的幅度值估计后，就可以按照从低频到高频（右向）和从高频到低频（左向）两个方向分别得到右向差分幅度谱和左向差分幅度谱。右向差分幅度谱的计算公式如下：

$$D_{right}(k)=\begin{cases}Y(k)-\alpha\cdot Y(k+1), & \text{若 } A'(k)>Y(k) \\ & \text{且 } A'(k+1)<Y(k+1) \\ \alpha\cdot Y(k)-Y(k+1), & \text{若 } A'(k)\le Y(k) \\ & \text{且 } A'(k+1)\ge Y(k+1) \\ Y(k)-Y(k+1), & \text{其他}\end{cases} \tag{2-9}$$

其中式（2-9）里的前两个条件分别对应图 2.3 中的（a）和（c）。

左向差分幅度谱的计算公式如下：

$$D_{left}(k)=\begin{cases}Y(k)-\alpha\cdot Y(k-1), & \text{若 } A'(k)<Y(k) \\ & \text{且} A'(k-1)<Y(k-1) \\ \alpha\cdot Y(k)-Y(k-1), & \text{若 } A'(k)\ge Y(k) \\ & \text{且} A'(k-1)\ge Y(k-1) \\ Y(k)-Y(k-1), & \text{其他}\end{cases} \tag{2-10}$$

其中式（2-10）里的前两个条件分别对应图 2.3 中的（b）和（d）。式（2-9）和式（2-10）中的 α 是一个经验值，在实验中设为 1.05。

（2）累积过程

通过预测差分运算后，本文使用累积运算来恢复干净的语音幅度谱。左向累积和右向累积运算如下：

$$Y'_{left}(k)=Y'_{left}(k-1)+D_{left}(k-1),\quad 1\le k\le N-1 \tag{2-11}$$

$$Y'_{right}(k)=Y'_{right}(k+1)+D_{right}(k+1),\quad 0\le k\le N-2 \tag{2-12}$$

在累积运算前，$Y'_{left}(0)$和$Y'_{right}(N-1)$的值设为 0。在得到恢复后的左向幅度谱和右向幅度谱之后，使用式（2-13）来得到最终的语音幅度谱。

$$Y'(k)=\frac{Y'_{left}(k)+Y'_{right}(k)}{2} \tag{2-13}$$

图 2.4 给出了原始语音、加噪语音、经过 NSS 后的语音以及经过 PDAS 后的语音。其中图 2.4(a)为原始干净语音，图 2.4(b)为混入信噪比为 0dB 的白噪声，图 2.4(c)为经过 NSS 处理后的语音，图 2.4(d)为经过 PDAS 处理后的语音。从图中可以看出，经 PDAS 处理后的语音相对于 NSS 来说，噪音的能量减少了很多。

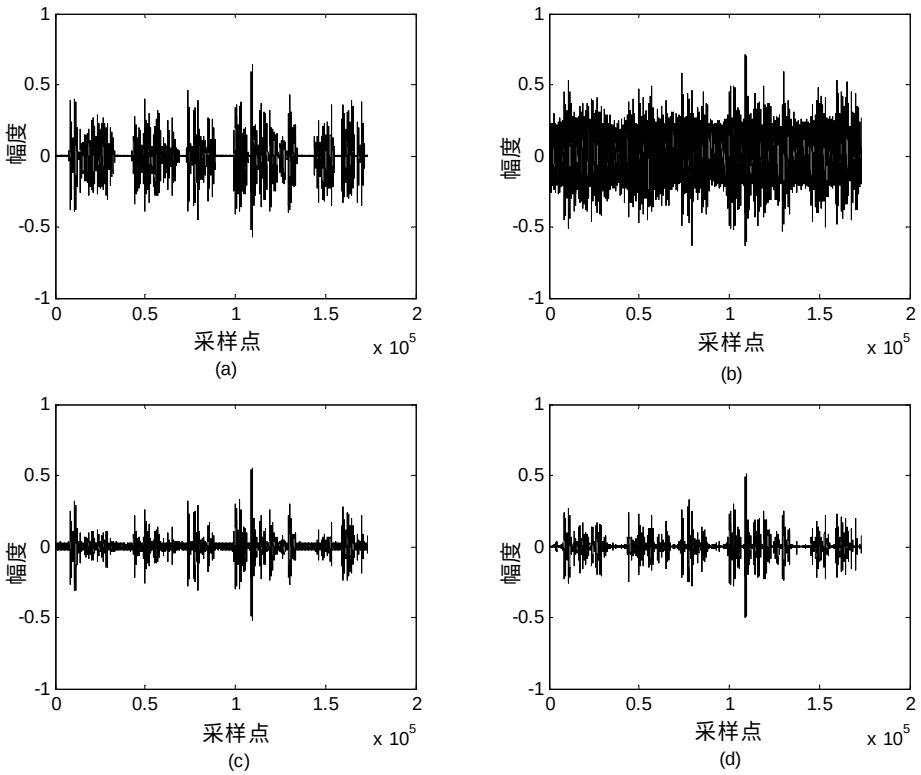

图 2.4 干净语音、含噪语音和恢复后的语音

2.3 实验结果与分析

2.3.1 实验设置和数据库

本实验中使用的系统都是基于 GMM-UBM 的说话人识别系统，

UBM 是用男女各 300 位说话人的语音数据训练得到的，总量为十个小时。训练 UBM 使用的说话人集合跟测试说话人集合没有重叠。本章实验中的每种系统使用的 UBM 都是在相同的数据集上训练得到的，并且每个系统的有效音检测（Voice Activity Detection，VAD）算法是相同的。

实验中每种系统的前端处理使用的帧长为 20 毫秒，帧移是 10 毫秒，预加重系数为 0.97，窗函数为哈明窗（Hamming Window），每帧语音使用的 FFT 大小为 256，截止频率为 200Hz～3800Hz，Mel 滤波器组的个数为 30。后面章节里所用到的语音特征提取参数均按上述规格处理，以后不再赘述。

本实验使用的数据库含有 522 个说话人（347 男、175 女）。语音格式为 8kHz 采样率，16 位精度，是固定电话信道下录制的语音。每个话者说 3 句话，其中 1 句用于训练，2 句用于测试。在干净测试语音中按照不同信噪比（SNR＝0，5，10，15，20dB）分别添加四种不同种类的噪音（White 噪音，Factory 噪音，Babble 噪音，F16 噪音）来得到含噪测试语音。噪音数据来自 NoiseEx-92 噪音数据库[101]。实验中训练语音的有效长度为 24 秒，测试语音的有效长度为 3 秒。实验中评测了三种系统：基于非线性谱减法（NSS）的系统；基于差分能量谱（DPS）的系统和基于预测差分幅度谱（PDAS）的系统，其中基于 NSS 的说话人识别系统作为 Baseline 系统。

2.3.2 实验结果与分析

实验一 White 噪音数据测试

White 噪音数据库测试结果见表 2.1。

表 2.1 中的粗体数值是本文提出算法（PDAS）得到的结果。“平均”指得是每种系统在不同信噪比的噪音下的平均识别率，最后一列给出 DPS 和 PDAS 相对于 NSS 在平均识别率下的错误率下降（Error Rate Reduction，ERR），本章后面实验表格与此类似。EER 的定义如下

$$EER = \frac{E_{old} - E_{new}}{E_{old}} \times 100\% \qquad (2\text{-}14)$$

其中 E_{new} 是新算法的错误率，E_{old} 是旧算法的错误率。

表 2.1 White 噪音数据上 NSS、DPS 和 PDAS 的测试结果

识别率(%) SNR(dB) / 系统	干净	20	15	10	5	0	平均(%)	ERR(%)
NSS	94.8	84.7	72.2	45.6	21.3	3.1	53.6	
DPS	94.6	89.7	82.0	65.5	41.2	19.7	65.5	26.6
PDAS	**95.8**	**93.1**	**88.5**	**66.9**	**42.2**	**21.5**	**68.0**	**31.0**

实验二 **Factory** 噪音数据库测试

Factory 噪音数据库测试结果见表 2.2。

表 2.2 Factory 噪音数据上 NSS、DPS 和 PDAS 的测试结果

识别率(%) SNR(dB) / 系统	干净	20	15	10	5	0	平均(%)	ERR(%)
NSS	94.8	93.3	91.2	87.6	75.1	35.8	79.6	
DPS	94.6	93.3	92.5	89.9	84.5	65.9	86.8	35.3
PDAS	**95.8**	**93.9**	**93.1**	**91.4**	**84.5**	**65.7**	**87.4**	**38.2**

实验三 **Babble** 噪音数据库测试

Babble 噪音数据库测试结果见表 2.3。

表 2.3 Babble 噪音数据上 NSS、DPS 和 PDAS 的测试结果

识别率(%) SNR(dB) / 系统	干净	20	15	10	5	0	平均(%)	ERR(%)
NSS	94.8	94.1	92.7	88.7	76.4	38.7	80.9	
DPS	94.6	93.1	89.1	81.0	63.2	34.3	75.9	-26.2
PDAS	**95.8**	**93.5**	**93.1**	**89.1**	**75.1**	**35.8**	**80.4**	**-2.6**

实验四 **F16** 噪音数据库测试

F16 噪音数据库测试结果见表 2.4。

表 2.4 F16 噪音数据上 NSS、DPS 和 PDAS 的测试结果

识别率(%) SNR(dB) / 系统	干净	20	15	10	5	0	平均(%)	ERR(%)
NSS	94.8	93.3	90.2	82.6	68.0	39.3	78.0	
DPS	94.6	93.3	91.2	87.6	75.1	55.4	82.9	22.3
PDAS	**95.8**	**93.7**	**92.5**	**88.3**	**71.5**	**51.2**	**82.2**	**19.1**

为了评价上述三种算法在不同噪音下的性能，表 2.5 给出三种算法在不同噪音下的平均性能和总的性能。

表 2.5 NSS、DPS 和 PDAS 的在四种噪音下的平均性能比较

识别率(%) Noise / 系统	White	Factory	Babble	F16	平均(%)	ERR(%)
NSS	53.6	79.6	80.9	78.0	73.0	
DPS	65.5	86.8	75.9	82.9	77.8	17.8
PDAS	**68.0**	**87.4**	**80.4**	**82.2**	**79.5**	**24.1**

2.3.3 讨论

从上面的实验可以看出，PDAS 能够较好的提高说话人识别系统的噪音鲁棒性，在四种噪音说话人辨识数据库下，平均错误率相对于 NSS 下降了 24.1%。在信噪比不小于 10dB 的情况下，PDAS 效果很好；但是在信噪比小于 10dB 的情况下：(1) PDAS 在 Babble 噪音下性能比 NSS 有所下降，这是因为 Babble 噪音是由背景集外说话人的语音组成的，在信噪比较低的情况下，这些语音与目标说话人的语音混在一起，使得 PDAS 在估计频谱的峰谷信息时产生了一定的错误；(2) PDAS 在 F16 噪音下性能比 DPS 有所下降，这是因为 F16 噪音谱的特点是在 600Hz-800Hz 和 2700Hz-2800Hz 上有着很强的能量，这会影响 PDAS 估计这些频率上峰谷信息的正确性，从而导致辨识率的下降。同样对 DPS 来说也存在类似的情况，因此 DPS 和 PDAS 在 F16 噪音下的平均性能相差不多。但对于 NSS 来说，由于 600Hz-800Hz 和 2700Hz-2800Hz 含有较多说话人的特定信息，通过对 F16 噪音谱求差，在去除噪音的同时，也去除了这些频率范围上的说话人信息，导致系统识别性能的下降。

2.4 小结

本章针对电话信道下说话人识别系统面临的背景噪音干扰问题，提出了基于预测差分幅度谱的特征提取算法，来尽可能恢复含噪语音中原始干净语音在频谱上的峰谷信息，以此减少测试语音与训练语音之间的不匹配。该算法主要特点如下：

第一，不需要对噪音做预估计；

第二，为了较好的估计当前频率点所处的峰谷位置，使用了正弦滤波器来预测（或估计）当前频率点右侧附近可能的幅度值。根据估计得到的幅度值来判断当前频率点所处的峰谷位置。

第三，在信噪比较高（SNR>10dB）的情况下，含噪语音谱受到噪音的影响相对较小，这时对频谱中峰谷信息的估计比较准确，采用加权差分函数和累积运算后，能够较好的减少测试语音和训练语音之间的不匹配。但是在信噪比较低的情况下，对频谱中峰谷信息的估计错误较多，影响了系统的识别性能。

第3章 说话人分割聚类研究

随着电视频道和广播电台的增多以及大容量存储设备的出现，越来越多的电视、广播语音被保留下来。如何快速便捷地从中查找到所需的音频信息（即音频内容检索），已逐渐成为人们关心的问题和研究的热点。音频信息中的内容大致可以分为：语音、音乐、环境音和静音四大类。其中对语音的检索称为说话人检测。所谓说话人检测，是指从一段多人语音里找出谁在说话，和在什么时候说话。一般来说，说话人分割和说话人聚类是说话人检测中的两个步骤。前者是指从多人语音中找到话者身份发生变化的时间点；后者是指从多人语音中找出话者的数目和每个话者在什么时候说话，即按照语音段话者的身份进行归类。

大多数情况下，语音流中话者的身份信息和话者的数目是未知的。而且国内外的研究者在进行说话人分割聚类时，通常假设语音流中不存在重叠语音的情况或者不关心重叠语音的情况（即多人同时发音的情况）。另外电话语音有其自身的特点，就是既含有部分较长的单人语音段，也含有部分较短的单人语音段。对其中较短的单人语音段的分割和身份归类（聚类）是影响说话人检测性能的一个主要因素，也是本章要处理的一个重点研究内容。

本章的内容安排如下：第一小节简单介绍一些常用的说话人分割聚类算法及研究现状；第二小节介绍本文提出的基于 UBM 的说话人分割聚类算法；第三小节给出实验结果和分析；最后一节是对本章内容的总结。

3.1 说话人分割聚类研究的现状

说话人分割聚类算法主要可以分为两大类：基于距离度量的分割聚类算法和基于模型搜索的分割聚类算法。前者是利用一定的距离度量准则来判断两段语音是属于同一个说话人还是属于不同的说话人；后者是利用得到的说话人模型来对原始多人语音按窗进行搜索，以便找出该话者发音的时间信息。近些年来，国内外研究人员加大了对说话人分割聚

类的研究力度，提出了不少具有实用价值的算法，下面本文将对这两类算法进行简单的综述。

3.1.1 基于距离度量的说话人分割聚类算法

这类算法主要是采用一定的距离度量准则来判断两段语音的接近程度，如果它们的距离小于预先设置的阈值，则判定这两段语音属于同一个说话人；否则属于不同的说话人。下面将简单介绍一些常用的基于距离度量的说话人分割聚类算法。

3.1.1.1 **BIC**

Chen 等人在 1998 年采用了 BIC（又被称为 Akaike or Rissanen 准则[102]）做为说话人分割的一种可分性度量方法。BIC 是一种基于模型复杂度（也就是模型参数）惩罚的最大似然准则。给定一段语音提取的特征序列 $F=\{f_1,f_2,\cdots,f_N\}$，它生成的模型为 M，$L(F|M)$是 F 在模型 M 上的似然得分。如果 m 是模型参数的个数，那么模型 M 上的 BIC 定义如下：

$$BIC(M)=\log L(F\mid M)-\lambda\frac{m}{2}\log N \qquad (3\text{-}1)$$

等号右边第一项表示数据与模型的匹配程度，第二项表示模型复杂度的惩罚分，其中 λ 是一个两项间的可调平衡参数。由于 BIC 不需要任何说话人的先验知识，就能够描述模型与数据的匹配程度，因此它常被用来做为说话人分割的一种可区分性度量方法。BIC 在说话人分割中的应用如下：

设 $F=\{f_1,f_2,\cdots,f_N\}$是一段语音提取出的特征序列，做如下两个假设：

H_0：如果该段语音属于同一个说话人 X，那么可以使用一个多维单高斯分布来描述该说话人，即$\{f_1,f_2,\cdots,f_N\}\sim N(\mu_X,\Sigma_X)$。

H_1：如果该段语音属于两个说话人 Y 和 Z，并在时刻 i 发生说话人转换，那么可以用两个多维单高斯分布来描述这两个说话人，也就是说$\{f_1,f_2,\cdots,f_i\}\sim N(\mu_Y,\Sigma_Y)$和$\{f_{i+1},f_{i+2},\cdots,f_N\}\sim N(\mu_Z,\Sigma_Z)$。

这里 μ 和 Σ 分别表示均值向量和协方差矩阵，也就是模型的参数。

那么 H_0 和 H_1 之间的 BIC 距离定义为：

$$R(i)=\frac{N}{2}\log|\Sigma_X|-\frac{N_Y}{2}\log|\Sigma_Y|-\frac{N_Z}{2}\log|\Sigma_Z| \tag{3-2}$$

其中，N 表示序列$\{f_1,f_2,\cdots,f_N\}$中特征的个数，N_Y 和 N_Z 分别表示序列$\{f_1,f_2,\cdots,f_i\}$和$\{f_{i+1},f_{i+2},\cdots,f_N\}$中特征的个数。一般来说，说话人转移发生在 $R(i)$出现极值的位置。

BIC 的另一种表示方式如下：

$$\Delta BIC(i)=-R(i)+\lambda P \tag{3-3}$$

其中 λ 是一个可调平衡参数，P 定义如下：

$$P=\frac{1}{2}\left(D+\frac{1}{2}D(D+1)\right)\times\log N \tag{3-4}$$

式（3-4）中 D 是特征的维数。如果$\Delta BIC(i)$小于 0，意味着该段语音属于两个说话人，分割位置在时刻 i；否则属于一个说话人。

基于 BIC 的说话人分割聚类算法得到了广泛的应用，在不同类型的数据库上都取得了较好的效果[103~105]。

3.1.1.2 KL 距离

给定两个变量 X 和 Y，它们的概率分布为 $P(X)$和 $P(Y)$，那么 X 和 Y 之间的 KL 距离定义如下：

$$KL(X,Y)=E_X\left[\left(\log P(X)-\log P(Y)\right)\right] \tag{3-5}$$

其中 $E_X[\cdot]$表示期望运算。

对于高斯分布来说，KL 距离可以表示如下：

$$\begin{aligned} KL(X,Y) &= \frac{1}{2}(\mu_Y-\mu_X)^T\left(\Sigma_X^{-1}+\Sigma_Y^{-1}\right)(\mu_Y-\mu_X) \\ &+\frac{1}{2}tr\left(\left(\Sigma_X^{1/2}\Sigma_Y^{-1/2}\right)\left(\Sigma_X^{1/2}\Sigma_Y^{-1/2}\right)^T\right) \\ &+\frac{1}{2}tr\left(\left(\Sigma_X^{-1/2}\Sigma_Y^{1/2}\right)\left(\Sigma_X^{-1/2}\Sigma_Y^{1/2}\right)^T\right)-D \end{aligned} \tag{3-6}$$

这里 $tr(\cdot)$ 表示矩阵的迹运算，D 为特征的维数。为了简化 KL 距离的计算，也可以采用如下的式子[106]：

$$\begin{aligned} KL(X,Y) &= \frac{1}{2}(\mu_Y-\mu_X)^T\left(\Sigma_X^{-1}+\Sigma_Y^{-1}\right)(\mu_Y-\mu_X) \\ &+\frac{1}{2}tr\left((\Sigma_X-\Sigma_Y)\left(\Sigma_Y^{-1}-\Sigma_X^{-1}\right)\right) \end{aligned} \tag{3-7}$$

等号右边第一项的值由均值和方差决定，第二项的值由方差决定。为了简化式（3-7），有的研究人员只使用第二项来计算 X 和 Y 的 KL 距离[106]，即

$$KL(X,Y)=\frac{1}{2}tr\left((\Sigma_X-\Sigma_Y)\left(\Sigma_Y^{-1}-\Sigma_X^{-1}\right)\right) \tag{3-8}$$

对于说话人分割来说，若 X 和 Y 表示两段短语音，那么如果 X 和 Y 的 KL 距离大于预先设定的阈值，则这两段语音属于不同的说话人；否则属于同一个说话人。

3.1.1.3 GLR 准则

给定一段特征序列 $F=\{f_1,f_2,\cdots,f_N\}$，做以下两个假设：

H_0：如果该段语音属于同一个说话人 X，那么可以使用一个多维单高斯分布来描述该说话人，即 $\{f_1,f_2,\cdots,f_N\}\sim N(\mu_X,\Sigma_X)$。

H_1：如果该段语音属于两个说话人 Y 和 Z，并在时刻 i 发生说话人转换，那么可以用两个多维单高斯分布来描述这两个说话人，即 $\{f_1,f_2,\cdots,f_i\}\sim N(\mu_Y,\Sigma_Y)$ 和 $\{f_{i+1},f_{i+2},\cdots,f_N\}\sim N(\mu_Z,\Sigma_Z)$。

则 H_0 和 H_1 之间的 GLR 距离定义如下：

$$R=\frac{L\left(F\mid N\left(\mu_X,\Sigma_X\right)\right)}{L\left(F_Y\mid N\left(\mu_Y,\Sigma_Y\right)\right)\cdot L\left(F_Z\mid N\left(\mu_Z,\Sigma_Z\right)\right)} \tag{3-9}$$

其中 F_Y 和 F_Z 分别表示特征序列 $\{f_1,f_2,\cdots,f_i\}$ 和 $\{f_{i+1},f_{i+2},\cdots,f_N\}$，$L(F|N(\mu_X,\Sigma_X))$表示 F 在高斯分布 $N(\mu_X,\Sigma_X)$上的似然得分。

在应用中，通常使用式（3-9）的 log 值来表示两段语音间的 GLR 距离，即：

$$d_R=-\log R \tag{3-10}$$

R的值越大（d_R 越小）表示 H_0 成立，也就是说两段语音属于同一个说话人；R的值越小（d_R 越大）表示 H_1 成立，也就是说两段语音属于不同的说话人。

GLR 准则不仅可用于说话人分割[107,108]，同时也可用于说话人确认，两处都取得了不错的效果[49]。

3.1.1.4 **DISTBIC**

如果多人语音中属于每个说话人的语音段都较长,那么 BIC 能够有较好的分割效果，但是对于每个说话人的语音段较短的情况（如对话交谈语音），其分割效果不是很好。考虑到 GLR、KL 距离等度量方法能够较好的处理短语音段，因此法国研究人员 P. Delacourt 等人提出了一种综合这些度量方法的分割算法：DISTBIC[48]。

DISTBIC 由初始分割和 BIC 细化两步组成。初始分割使用的是 GLR、KL 距离和一些描述两段语音相似程度的准则[109]，按照这些度量准则计算出语音段的距离序列，并对序列中的极值进行判断，来确定该极值对应的时间点是否为一个说话人切换点；BIC 细化则是在初始分割的基础上，用 BIC 来判断初始分割中相邻的两个语音段是否应该合并。

3.1.1.5 交叉似然比

交叉似然比（Cross Likelihood Ratio，CLR）[110]的定义如下：

$$d_{clr}(X,Y)=\frac{L(Y|\lambda(W))}{L(Y|\lambda(X))}\cdot\frac{L(X|\lambda(W))}{L(X|\lambda(Y))} \tag{3-11}$$

其中，X 和 Y 是分割后的语音提取的特征序列，$\lambda(W)$是背景模型，$\lambda(X)$是从 X 上得到的模型，$\lambda(Y)$是从 Y 上得到的模型，$L(\cdot)$表示似然分运算。

Meignier 等人[111]提出了两种改进的交叉似然比，分别定义如下：

$$d_{clr}(X,Y)=\frac{L(\overline{Y}|\lambda(X))+L(\overline{X}|\lambda(Y))}{L(Y|\lambda(X))\cdot L(X|\lambda(Y))} \tag{3-12}$$

$$d_{clr}(X,Y)=\frac{L(Y|\lambda(\overline{X}))+L(X|\lambda(\overline{Y}))}{L(Y|\lambda(X))\cdot L(X|\lambda(Y))} \tag{3-13}$$

其中 $\overline{X}$ 和 $\overline{Y}$ 分别表示除 X 和 Y 之外的其他语音提取的特征序列，也就是“非”的意思。

上述准则既可以用于说话人分割，也可以用于说话人聚类。一般来说，基于距离度量的说话人聚类算法大多采用自底向上（Bottom-Up）的聚类方式[42]，图 3.1 给出第一步迭代的示意图：

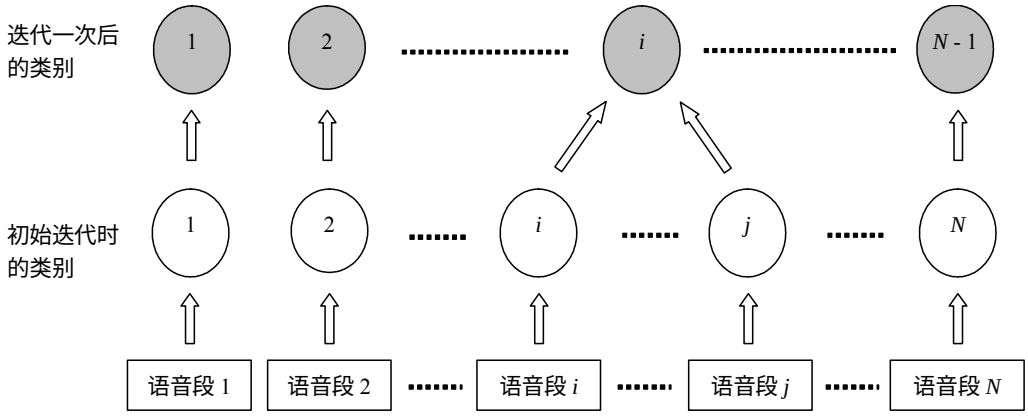

图 3.1 自底向上一步聚类示意图

自底向上的说话人聚类算法一开始将分割后的 N 个语音段各自生成 N 个模型，按照选定的距离度量准则计算任意两个模型之间的距离，并将距离最近的两段语音合并，生成新的模型，重复上述步骤直到满足

收敛条件。收敛条件一般是达到所需的聚类数目，或者是模型间最短距离大于给定的阈值。

基于距离度量的说话人分割算法的一个难点在于阈值的确定，因为：(1) 在实际应用时，语音信号千差万别，不同说话人的语音段求得的距离值差异很大，有的很明显，有的则不明显；(2) 即使两段语音属于同一个说话人，由于数据长度的不同，话者情感的不同等原因，也会使得计算出的距离值差异较大。因此，基于距离度量的说话人分割算法有一定的应用局限性，对于不同的应用场景或说话人，需要设定不同的阈值。

3.1.2 基于模型搜索的说话人分割聚类算法

基于模型搜索的说话人聚类是根据从语音段中生成的特定说话人模型来对原始语音进行搜索，通过不断的更新特定说话人的模型并对原始语音进行重搜索来完成说话人检测任务。下面简单介绍几个基于模型搜索的说话人分割聚类系统。

3.1.2.1 **LIA** 系统

LIA 系统[112]是建立在 HMM 上的，HMM 中的每个状态代表一个说话人，状态之间的转移代表说话人的切换。具体步骤如下：

(1) 初始化：用整段语音训练模型 S_0 做为一个状态建立单状态的 HMM。

(2) 添加新模型：用长度为 3 秒的滑动语音窗对 S_0 计算似然分，取似然得分最大的那窗语音训练新的模型，标记为 S_n（n 是迭代的次数），并把该模型作为一个新的状态添加到以前的 HMM 中去。

(3) 更新模型：首先按照当前的分割结果，用属于同一个说话人的语音更新该人原先的模型。接着使用 Viterbi 解码得到新的分割结果，也就得到了分割后每段语音的说话人身份。重复模型更新和解码过程，直到相邻两次解码结果不变。

(4) 收敛准则：迭代结束条件取决于相邻两次 Viterbi 路径得分的差和模型 S_n 所对应的语音段的个数（如果 S_n 只对应一段语音，那么迭代结束，取前一次迭代的结果）。

3.1.2.2 **CLIPS** 系统

CLIPS 系统[43]的基本流程如下：

（1）初始分割：按 1.5 秒的固定窗长用 BIC 进行初始分割。

（2）聚类：首先用整段语音训练一个 32 混合的 GMM 背景模型，协方差矩阵为对角矩阵。用分割出的每段语音在此背景模型上自适应出相应的模型，按照自底向上的方法进行聚类，直到聚为 N 类（对于对话交谈语音，N 一般设为 2）。

（3）重分割：用前面得到的 N 个模型对固定窗长为 0.8 秒的语音计算似然分，按照最大得分来判断该窗语音属于哪个说话人。

3.1.2.3 **ELISA** 系统

ELISA 系统[43]是建立在 LIA 系统和 CLIPS 系统之上的，它有两种结合方式：杂交（Hybridization）和融合（Fusion）。所谓杂交是指将一种系统的结果作为另一种系统的初始输入；所谓融合是指将两种系统的结果“求交”，分割结果一致的保留不变，对于分割结果不同的语音段，采用任一系统进行重新分割。

基于模型搜索的算法在分割聚类性能上一般好于基于距离度量的算法，但是它的时间花销往往是后者的许多倍，而且它有一个需要注意的地方，就是在初始模型训练时，如果选择的语音段不恰当（即包含有多个人的语音），就会使得用于搜索的模型不正确或不精确，导致最后分割聚类的结果不好。

3.1.3 评测指标

说话人分割算法常用的评测指标有两个，一个是误警率（False Alarm Rate，FAR），另一个是漏检率（Miss Detection Rate，MDR）。所谓误警是指给出的分割点实际上并不存在，也就是说分割点左右相邻的两段语音是属于同一个说话人；所谓漏检是指没有给出实际存在的说话人分割点，也就是说分割后的语音段里含有多个说话人。

FAR 的定义如下：

$$FAR = \frac{\text{误警的个数}}{\text{实际分割点的个数}+\text{误警的个数}} \times 100\% \qquad (3\text{-}14)$$

MDR 的定义如下：

$$MDR = \frac{\text{漏检的个数}}{\text{实际分割点的个数}} \times 100\% \qquad (3\text{-}15)$$

一般来说，FAR 越低，MDR 就相对越高，反之亦然。对一个说话人识别系统来说，MDR 的危害性要远大于 FAR，但不是说 MDR 越低越好，这是因为 MDR 越低，FAR 相应就越高，分割后得到的语音段平均长度也就越短，不利于后面的说话人聚类和识别处理。在本文中，采用的说话人分割算法评价准则为上述两个错误率。

由于实验数据库选用的是 NIST 2002 Switchboard 说话人分割数据库，因此在本文中使用了 NIST 提供的 Perl 评测脚本（版本号 v07）[113] 来评价分割聚类算法的性能，该评测脚本有五个评测指标：

- PMissSeg：丢失的语音段（即误标为非语音的语音段）；
- PFASeg：接受的非语音段（即误标为语音的非语音段）；
- PMissSpkr：漏掉的特定说话人语音段（比如原始语音有两个说话人 *A* 和 *B*，分类结果为 *C* 和 *D*，但是 *C* 的大部分语音段属于 *A*，*D* 的所有语音段也属于 *A* 并且 *C* 中属于 *A* 的语音段总长度大于 *D*，那么 *C* 对应于 *A*，*D* 无对应的说话人，属于 *D* 的语音段就是 PMissSpkr）；
- PFASpkr：错误接受的特定说话人语音段（前提假设同 PMissSpkr，由于 *C* 对应于 *A*，*D* 无对应的说话人，那么属于 *C* 但不属于 *A* 的语音段就是 PFASpkr）；
- PErrSpkr：分类错误的说话人语音段（比如原始语音有两个说话人 *A* 和 *B*，分类结果为 *C* 和 *D*，并且 *C* 对应于 *A*、*D* 对应于 *B*，那么属于 *C* 但不属于 *A* 的语音段和属于 *D* 但不属于 *B* 的语音段就是 PErrSpkr）。

上述对于说话人分割聚类来说比较重要的指标是后三个。该脚本还

给出了一个总的聚类错误率 C_{seg}，定义如下：

$$\begin{aligned} C_{seg} = & \left(C_{MissSeg} \cdot P_{MissSeg} + C_{FASeg} \cdot P_{FASeg}\right) \\ & + \left(C_{MissSpkr} \cdot P_{MissSpkr} + C_{FASpkr} \cdot P_{FASpkr}\right) \\ & + C_{ErrSpkr} \cdot P_{ErrSpkr} \end{aligned} \tag{3-16}$$

其中 C_X 表示错误率（X 表示上面五种错误类型），P_X 表示对应的错误权重（X 表示上面五种错误类型）。

除了上述评价指标外，研究人员常用的有：

（1）纯语音比例（Pure Rate）：纯语音段的定义是只含有一个说话人语音的语音段，含有噪音或者多个说话人语音的语音段则不是纯语音段。纯语音比例是指纯语音段总长度占全部语音长度的比例，定义如下：

$$Pure = \frac{\text{纯语音段总长度}}{\text{实际语音段的长度}} \times 100\% \tag{3-17}$$

（2）召回率（Recall Rate）：指检测到的正确的切换点在所有实际的切换点中所占的比例。

（3）精确率（Precision Rate）：指检测到的正确的切换点在检测到的所有切换点中所占的比例。

3.2 基于UBM的说话人分割聚类算法

对多人语音进行分割聚类处理的一个常用思路是先对多人语音进行分割处理，尽可能的找出话者发生转换的时间点；然后对分割的结果进行聚类处理，即按照分割后语音段话者的身份进行归类。本文提出的基于 UBM 的说话人分割聚类算法也是按照这一思路进行的，可以分为三个阶段：初始分割、聚类和重分割。由于电话交谈语音中含有较多的短语音段，对这些短语音段处理的好坏将影响到说话人分割算法的性能。而常用的说话人分割算法一般是基于语音窗分析的，并要求说话人一次发音的时间不能太短，这是因为太短的语音段会使得单个语音窗内含有较多的说话人转换点，影响分割算法的性能。考虑到 UBM 代表了大多数说话人的发音特性，是一个高精度的模型，并且语音段在 UBM

上的似然分差异也能够在一定程度上反映出它们之间在声学分布上的不同，因此，在分割阶段，本文将语音段在 UBM 上的似然比分作为一种距离度量准则，用以找出交谈语音中可能的说话人转换点。在聚类阶段，由于电话交谈语音一般只含有两个说话人，因此本文对所研究的问题做了简化，假定多人语音中只含有两个说话人。为了对语音中含有的两个说话人建立正确的说话人模型，并在此基础上对分割后的语音段按照话者的身份进行归类，本文提出了一种基于模型间分数差的聚类方法。该算法将一段语音在两个模型上的分数差作为该语音段属于其中某个模型的“概率分”，并根据“概率分”的大小，选择得分较大的语音段用于训练或更新特定说话人的模型，以保证训练用语音段的话者身份尽可能一致。由于用于训练初始说话人模型的语音段一般较短，因此采用了在小混合的 UBM 上自适应得到说话人模型的方式，来保证说话人模型的精度。在对语音段有了一个初步的归类结果后，用于训练说话人模型的语音段相对较多，可以使用大混合的 UBM，来提高说话人模型的精度，并修正上次的聚类结果。由于在聚类时，没有改变分割的结果，这就使得在分割时产生的漏检错误不能得到一定的消除，因此，本文进行了重分割处理，利用聚类时得到的说话人模型，对多人语音按窗进行重新分割。

在介绍本文提出的算法之前，我们先对实验中系统的参数设置和用到的数据库进行说明，以便对算法有一个更好的说明。

3.2.1 实验设置和数据库

由于实验室只购买了 NIST 2002 年的说话人分割聚类数据库，而且 NIST 组织的说话人识别评测也不包含对说话人分割聚类的评测，因此在本章的实验中使用的数据库是 NIST 2002 年的 Switchboard 分割聚类数据库。这是一个双人电话交谈语音数据库，共有 199 个语音文件，每个语音文件含 2 分钟的语音数据（包含有同性别和不同性别的话者的英文对话）。从 NIST 提供的说话人转换点标注文件来看，每个文件平均有 64 个转换点，即说话人一次发音的平均时间为 1.9 秒。

实验中使用的 UBM 是用 NIST 2002 年提供的电话信道单人测试库中的语音数据经 EM 算法训练得到的。该 UBM 共有 1024 个混合，并采

用了树型结构[114]。实验中使用的特征是 16 维 PDASCC（见第 2.2 节）和 16 维一阶差分系数，并经过倒谱均值减处理。

在统计分割结果的 FAR 和 MDR 时，本文给定了一个 0.2 秒的容错范围，也就是说如果得到的分割点与实际分割点的距离小于 0.2 秒的话，就算一次命中，否则判为误分割或者漏分割。

3.2.2 初始分割

要判断两段语音是否属于同一个说话人，一般来说，需要有一种好的距离（差异）度量准则。该准则应该既能很好的反映出不同说话人之间的差异，又能使得同一说话人自身的差异不明显。假设我们能够得到说话人 A 精度较高的模型 S_A，那么如果两段语音都属于说话人 A，则这两段语音在 S_A 上的得分差异一般较小；如果两段语音属于不同的说话人，则这两段语音在 S_A 上的得分差异一般比较大。并且由于模型 S_A 的精度较高，对语音段的长度要求也就相应较低。然而电话交谈语音中话者的身份信息是预先未知的，并且一般也不可能事先得到待检测说话人的模型。考虑到 UBM 较好的覆盖了说话人的声学发音分布，而且两段语音在 UBM 上的得分差也能够在一定程度上反映出了它们在声学分布上的差异，因此本文使用 UBM 来代替上面的特定说话人模型 S_A，并用 UBM 上两段语音的对数似然比分（Log-Likelihood Ratio Score，LLRS）来作为一种说话人分割的可区分性度量准则。

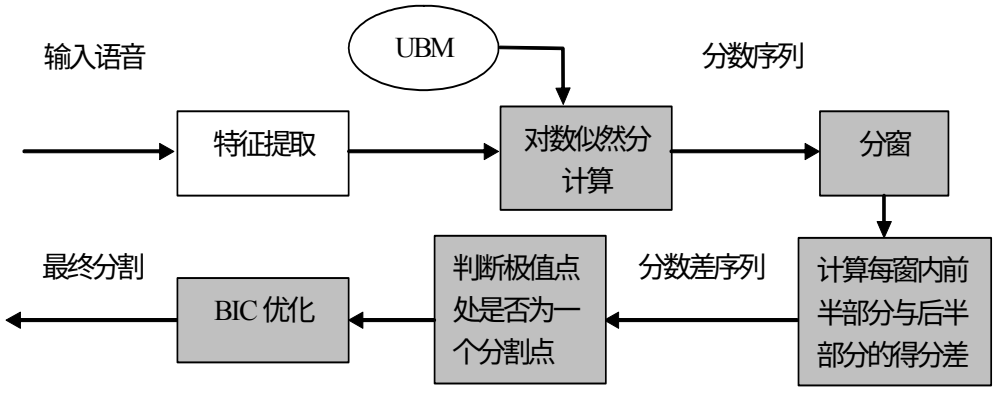

图 3.2 基于 UBM 上 LLRS 的说话人分割算法基本流程

基于 UBM 上对数似然比分的说话人分割算法包括五个步骤，如图

3.2 中阴影部分所示，这五个步骤是：

（1）对数似然分计算：对输入语音提取的每帧特征计算它在 UBM 上的对数似然分；

（2）分窗：将特征序列按照一定的窗长和窗移进行分窗，以便对每窗的中心进行说话人转换点的判断；

（3）对数似然比分计算：在中心位置将每窗特征分为两部分，计算出这两部分特征之间的对数似然比分；

（4）分割点判断：判断得到的对数似然比分序列中的极值点是否为一个说话人转换点；

（5）BIC 优化：降低在第（4）步中产生的误警率（即误判的分割点）。

下面详细介绍这五个步骤。

3.2.2.1 对数似然分计算

给定一个 D 维的特征向量 X 和 UBM，那么 X 在 UBM 上的似然函数定义如下：

$$p(X \mid \mathrm{UBM}) = \sum_{i=1}^{M} w_i g_i(X) \tag{3-18}$$

其中，M 是 UBM 中高斯混合的个数，w_i 是第 i 个高斯混合的权重，并满足：

$$\sum_{i=1}^{M} w_i = 1 \tag{3-19}$$

这里 $g_i(\cdot)$ 是期望为 μ_i，协方差矩阵为 Σ_i 的高斯混合的概率密度函数：

$$g_i(X) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)\right\} \tag{3-20}$$

一般来说，一帧语音在 UBM 上的似然分只统计最大的前 N 个混合

（一般 N 取 4 或 5）的得分，对于 1024 混合的 UBM 来说，要先计算出 1024 个混合的得分，然后按从大到小的选取前 N 个得分，这样计算的速度比较慢。为了节省计算混合得分的时间花销，在实验中采用了熊振宇提出的树型结构的 UBM[114]，这样一帧语音在 1024 混合的 UBM 上只需计算 28 个高斯混合的得分，然后从中选取前 N 个得分，较大的节省了计算混合得分的时间花销。如无特别说明，后面实验里用到的 1024 混合的 UBM 均是基于此树型结构的。

3.2.2.2 分窗

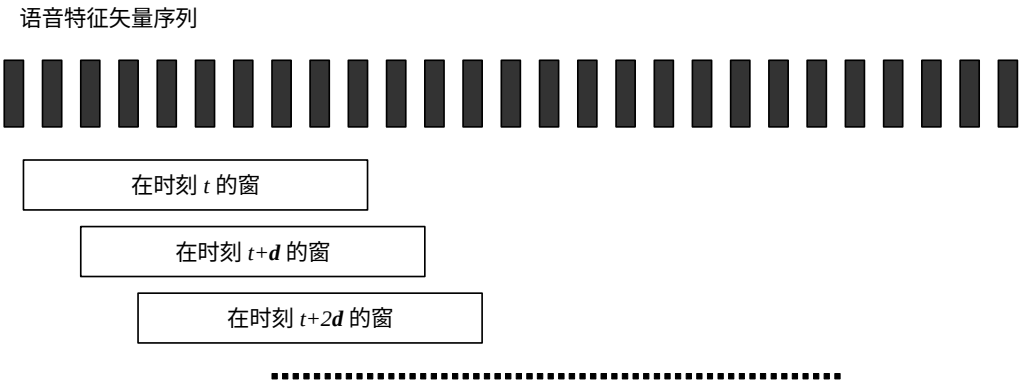

图 3.3 分窗示意图

分窗的示意过程可以参看图 3.3，图中 d 为窗移，即分割的精度。t 为每窗的中心位置所对应的时刻。

分窗跟声学特征提取中的分帧类似，也有窗长和窗移两个参数。一般来说，窗长不宜太长，否则会使窗内含有多个说话人转换点。通常窗长取 2 秒。窗移表示的是分割的精度，窗移越小，分割精度越高，分割耗时也就越大；窗移越大，分割精度就低，分割耗时也就越小。通常窗移取为 0.1 秒。根据待分割的语音类型不同，窗长和窗移也不同。对于广播或电视语音来说，由于说话人一次发音平均时间较长，因此窗长和窗移可以适当大一些；而对于交谈语音或对话语音来说，语音中含有较多的短发音，因此窗长和窗移一般略小些。

3.2.2.3 对数似然比分计算

在中心位置将一窗语音分为前后两个半窗，做如下假设：

H_0：如果该窗语音属于同一个说话人，那么前后两个半窗的语音在 UBM 上的对数似然分相差不大。

H_1：如果该窗语音属于两个说话人，那么前后两个半窗的语音在 UBM 上的对数似然分相差较大。

前后两个半窗的语音在 UBM 上的对数似然比分定义如下：

$$\Delta S(i) = abs\left(L(X_1 | UBM) - L(X_2 | UBM)\right) \qquad (3\text{-}21)$$

其中，$\Delta S(i)$是 i 时刻语音窗的对数似然比分，X_1 对应于前半窗语音，X_2 对应于后半窗语音，$L(X_1|UBM)$和 $L(X_2|UBM)$分别是前后两个半窗的语音在 UBM 上的似然得分。由于同一窗内的语音可能含有静音、背景噪音、音乐等等非说话人的声音，如果使用半窗内所有的语音来计算 $L(X_1|UBM)$和 $L(X_2|UBM)$的话，将会引入一定的误差。考虑到微软亚洲研究院提出的基于 UBM 的说话人实时分割算法[47]中，根据每帧语音在 UBM 上的得分大小，能够较好的将语音帧划分为可靠说话人语音帧、可疑说话人语音帧和非说话人语音帧，因此，在一窗语音内，本文使用了似然得分按从大到小的顺序排在前面的语音（即是说话人语音的可能性比较大）来计算 $L(X_1|UBM)$和 $L(X_2|UBM)$，以此减小非语音部分带来的影响。

3.2.2.4 分割点判断

通过计算每窗的对数似然比分，就可以得到一个由该分数组成的序列$\{\Delta S(i)\}$。该序列上的每个极大值点都有可能是一个说话人分割点，可以按照下面的方法来判断一个极大值点是否为一个说话人分割点。

假定该分数序列服从高斯分布，那么它的方差表示了数据的离散程度。由于前面已经假定每段语音只含有两个说话人，因此可以使用分数序列的方差来确定最后的分割点判决阈值。给定分数序列上的一个极大值点和它的左右相邻的两个极小值点，如果极大值点与其相邻两个极小值点对应的数值之差都大于给定的阈值，那么该极大值点就是一个可能

的分割点；否则，该极大值点不是分割点。图 3.4 给出了一段对数似然比分序列上三个可能的分割点：*A*，*B* 和 *C*。由于 *C* 点与其相邻的极小值点对应的数值之差都比较大，因此 *C* 点相对于 *A* 和 *B* 来说，成为分割点的可能性更大。算法中的判决阈值取为 $\alpha\cdot\sigma$，其中 α 是一个可调节的参数，σ 是在对数似然比分序列上求出的标准方差，公式如下：

$$Pos(\max)=\begin{cases}\text{分割点，若}\left|\max-\min_l\right|>\alpha\cdot\sigma \\ \qquad\quad\text{且}\left|\max-\min_r\right|>\alpha\cdot\sigma \\ \text{不是分割点，}\qquad\quad\text{否则}\end{cases} \tag{3-22}$$

其中 max 是序列 $\{\Delta S(i)\}$ 中的一个极大值，$\min_l$ 和 $\min_r$ 分别是 max 最近的左、右两个极小值。$Pos(\max)$ 表示 max 所对应的时间点。

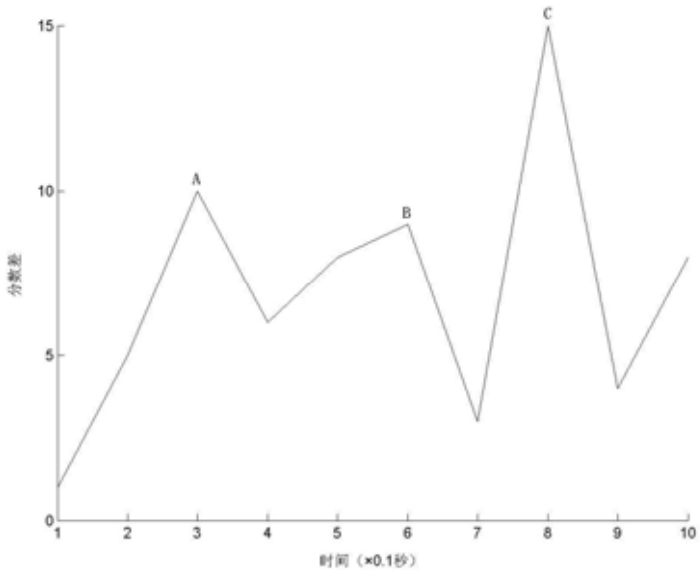

图 3.4 对数似然比分序列上的三个极值点

3.2.2.5 BIC 优化

通过步骤 3.2.2.4 后，我们得到了一组分割点，这些分割点中有的是正确的，有的是错误的。为了减少误判的分割点（即误警率），本文使用了 BIC 来判断一个分割点左右相邻的两段语音是否属于同一个说话人。如果两段语音属于同一个说话人，则合并为一段语音，并将该分

割点去除；如果两段语音不属于同一个说话人，则保留该分割点。通过 BIC 优化后，误警率能够得到一定的降低，但是漏检率并不能得到降低，反而会稍有提高。对漏检分割点的处理，将放在重分割阶段进行。在 BIC 优化中，使用的 BIC 距离的计算公式为式（3-3），其中 λ 在实验里取值为 1。

3.2.2.6 初始分割实验结果比较分析

在基于 UBM 上 LLRS 的初始分割中，使用的窗长为 2s，窗移为 0.1s。表 3.1 给出了分割点判断中判决阈值 α 取值对最后分割性能的影响。

表 3.1 不同 α 取值下的分割结果

α	**FAR(%)**	**MDR(%)**
1.0	21.0	28.9
0.8	23.1	27.5
0.5	25.2	19.0
0.3	29.7	18.8
0.1	30.1	18.5

从表中可以看出，随着 α 取值的减小，FAR 逐渐变大，MDR 逐渐变小。这是因为 α 的取值大小决定分割点判断的严格程度，α 取值越大，判断越严格，因此 FAR 就越小，MDR 就越大。而当 α 取值过小时，由于 FAR 比较高，使得分割后的语音段有较多的短语音段，导致后面的 BIC 优化效果不好。由于 α 取值为 0.5 时，算法的两错误率之和最低，因此在分割算法中 α 取值为 0.5。

同时本文还比较了基于 BIC、GLR、DISTBIC 和 LLRS 的四种说话人分割算法的性能，见表 3.2，其中 BIC 算法基于式（3-3）、GLR 算法基于式（3-10）。四种算法使用的窗长都是 2s，窗移为 0.1s。表 3.2 中还给出了在分割时未使用 BIC 优化的结果，记为“ LLRS+No BIC ”。

从表 3.2 中可以看出，基于 UBM 上 LLRS 的分割算法取得了较好的分割性能，相比于 DISTBIC 来说，总错误率（FAR＋MDR）相对下降了 13.5%。这是由于使用了 UBM 这一代表说话人的发音共性的先验

知识，使得从相邻两段语音上计算得到的“距离”能够比较好的反映出不同说话人的发音差异；经过 BIC 优化后，FAR 得到了进一步的降低，而 MDR 的提高不是很大。

表 3.2 BIC、GLR、DISTBIC 和 LLRS 的分割结果比较

算法	FAR(%)	MDR(%)
BIC	25.2	35.6
GLR	33.2	19.5
DISTBIC	30.8	20.3
LLRS+No BIC	29.3	18.9
LLRS	**25.2**	**19.0**

3.2.3 聚类

在初始分割后，我们可以得到许多只含有单个说话人的语音段，但是还不能知道哪几段语音是属于同一个说话人的，这时就需要采用聚类算法将语音段按照其话者身份进行归类。由于多人语音中话者的身份是未知的，所以这类算法属于无监督的聚类算法。一般来说，无监督的说话人聚类算法的好坏与初始聚类结果有较大的关系，而初始聚类常常会遇到两个问题：一个是由于语音段较短导致的说话人模型精度不足；另一个是如何选择生成说话人模型的初始语音段。

对于短语音导致的说话人模型精度不足的问题，可以采用从 UBM 上自适应出说话人模型的方式，来解决语音长度不够的问题。这样，短语音中含有的特定说话人发音分布可由自身来描述，短语音中未含有的发音，可以用 UBM 中的混合来代替，这样就保证了模型的精度。论文使用的自适应算法为 MAP 自适应算法，根据训练语音改变 UBM 的混合均值来得到特定说话人的模型。均值的改变采用如下公式：

$$\mu_k^S = \frac{\tau \cdot \mu_k^{UBM} + \sum_{t=1}^{T} c_{kt} \cdot o_t}{\tau + \sum_{t=1}^{T} c_{kt}} \tag{3-23}$$

其中 o_t 表示第 t 帧语音特征，T 为训练语音的帧数，k 为 GMM 的第 k 个混合，c_{kt} 为：

$$c_{kt} = \frac{w_k \cdot g_k(o_t)}{\sum_{i=1}^{M} w_i \cdot g_i(o_t)} \tag{3-24}$$

其中 $g_k(o_t)$为第 k 个高斯混合的概率密度函数，可以参看式（3-20），w_k 是 UBM 中第 k 个混合的权重，M 是 UBM 中高斯混合的个数。式（3-23）中的 τ 是模型先验分布的一个重要参数，它控制着自适应对先验信息 μ_k^{UBM} 的依赖程度。如果 τ 越大，自适应后的说话人模型参数越接近于 UBM 的参数；如果 τ 越小，则说话人模型的参数主要由训练语音决定。在训练数据有限的情况下，一般采用较大的 τ 值，在本文的实验里 τ 设为 16。

对于如何选择生成说话人模型的初始语音段的问题，若能够给出每个语音段属于某个说话人的“概率”度量，就可以从这些语音段中选择“概率”较大的语音段用于训练该说话人的模型。由于前面假设电话交谈语音中只含有两个说话人，因此本文使用了语音段在两个说话人模型上的得分差来作为该语音属于某个说话人的“概率”度量。给定两个模型 S_1 和 S_2 以及一段语音 X，计算 X 在两个模型上的对数似然分，分别记为 $L(X|S_1)$和 $L(X|S_2)$。若 $L(X|S_1)$-$L(X|S_2)$>0，则说明 X 属于 S_1 的可能性大于 S_2。对于另一段语音 Y 来说，它在两个模型上的得分为 $L(Y|S_1)$和 $L(Y|S_2)$。同样的，由于 $L(Y|S_1)$-$L(Y|S_2)$>0，因此 Y 属于 S_1 的可能性也大于 S_2。要判断 X 和 Y 中哪一个更有可能属于模型 S_1，可以比较 $L(X|S_1)$-$L(X|S_2)$和 $L(Y|S_1)$-$L(Y|S_2)$的大小，若前者大于后者，则说明 X 属于模型 S_1 的可能性大于 Y。这样，就使用 X 来对模型 S_1 进行更新，否则使用 Y。

本文提出的聚类算法包括两个步骤：初始聚类和迭代归类，如图 3.5 所示。在初始聚类时，由于分割后的每个语音段类别未知，因此在训练说话人模型时，用于训练的语音不是很多，因此采用了较小混合的 UBM 以保证模型的精度，在实验中使用的是 16 混合的 UBM。在得到初始聚类结果后，用于训练说话人模型的语音相对较多，因此采用了较大混合的 UBM，以保证语音段归类的准确性，在实验中使用的是 1024 混合的 UBM。

下面将分别介绍这两个步骤。

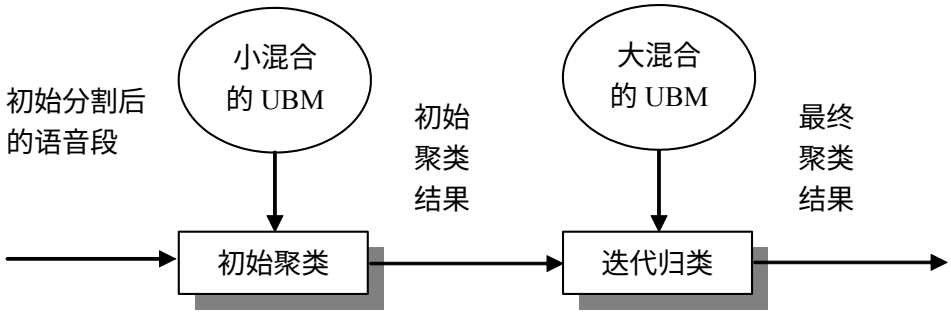

图 3.5 基于说话人模型间分数差的说话人聚类算法示意图

3.2.3.1 初始聚类

由于前面假定电话交谈语音只含有两个说话人，因此初始聚类的基本想法是首先从初始分割后的语音段中找出属于每个说话人“概率”最大的两段语音，并用这两段语音来训练说话人模型；然后根据得到的模型对剩余的语音段进行归类。初始聚类用到的 UBM 含有 16 个混合，记为 UBM1，具体描述如下：

1. 首先根据整段语音在 UBM1 上自适应得到模型 S_0，分割后的语音段标记为“未归类”状态；

2. 用分割后的每个语音段在 S_0 上求似然分，取得分最大并且长度大于 2 秒的语音段作为第一个说话人的模型训练语音，从 UBM1 上自适应得到该说话人的模型 S_1，并将该段语音标记为“已归类”；

3. 用“未归类”的语音段对 S_0 和 S_1 求似然分，并取它们之间的分数差（ΔS）最大并且长度大于 2 秒的语音段作为第二个说话人的模型训

练语音，从 UBM1 上自适应的到该说话人的模型 S_2，并将该段语音标记为“已归类”。ΔS 的计算公式如下：

$$\Delta S = L(X \mid S_0) - L(X \mid S_1) \tag{3-25}$$

其中 X 是用于计算似然分的语音段，ΔS 反映出该语音段与模型 S_1 的差距，差越大，属于 S_1 的可能性就越小；

4. 用模型 S_1 和 S_2，分别对“未归类”的语音段求似然分比 ΔS_{12} 和 ΔS_{21}，计算公式如下：

$$\Delta S_{12} = L(X \mid S_1) - L(X \mid S_2) \tag{3-26}$$

$$\Delta S_{21} = L(X \mid S_2) - L(X \mid S_1) \tag{3-27}$$

每次挑选：(1) ΔS_{12} 最大且长度大于 1 秒的语音段用于更新模型 S_1，标记该段语音的类别并记为“已归类”；(2) ΔS_{21} 最大且长度大于 1 秒的语音段用于更新模型 S_2，标记该段语音的类别并记为“已归类”。重复步骤 4 直到所有的语音都处理完毕，或不存在大于 1 秒的“未归类”语音段；

5. 最后用得到的模型 S_1 对属于 S_1 的语音段计算 ΔS_{12} 并保留；用模型 S_2 对属于 S_2 的语音段计算 ΔS_{21} 并保留。

3.2.3.2. 迭代归类

由于初始聚类使用的 UBM 精度不高，使得聚类的结果不是很好。为了进一步降低聚类错误，在初始聚类的基础上，进行迭代归类。这时，用于训练说话人的语音相对较长，因而可以使用大混合的 UBM 来得到精度较高的说话人模型；同时为了避免初始聚类中错误归类的语音段对模型训练产生不好的影响，在训练说话人模型时，仅使用“概率”较高的语音段。迭代归类使用了 1024 混合的 UBM，记为 UBM2，具体描述如下：

1. 对属于模型 S_1 的语音段的 ΔS_{12} 按从大到小进行排序，取数值靠前的一半在 UBM2 上重新训练模型 S_1；

2. 对属于模型 S_2 的语音段的ΔS_{21} 按从大到小进行排序，取数值靠前的一半在 UBM2 上重新训练模型 S_2；

3. 用新得到的模型 S_1 和 S_2 对所有语音段计算ΔS_{12}，如果ΔS_{12} 大于 0，则该语音段属于模型 S_1，否则属于模型 S_2。同时用模型 S_1 对属于 S_1 的语音段计算ΔS_{12} 并保留；用模型 S_2 对属于 S_2 的语音段计算ΔS_{21} 并保留；

4. 重复上述步骤直到本次聚类结果与上次相同或者达到最大迭代次数（实验里最大迭代次数取 4）。

由于说话人自身发音的差异，会使得该人不同的语音段的得分变化较大，影响ΔS_{12} 和ΔS_{21} 的数值大小，从而影响聚类的结果。为了在一定程度上减少说话人自身差异对语音段得分带来的影响，本文使用了 Dnorm 算法。Dnorm 是 Ben 等人[85]在 2002 年提出来的，利用用户模型 λ 和背景模型 $\overline{\lambda}$ 之间的 KL 距离来对用户模型 λ 的得分进行归一化处理的方法。Dnorm 的一个优点就是不需要额外的说话人数据，它的定义如下：

$$L'_{\lambda}(X)=\frac{L_{\lambda}(X)}{KL(\lambda,\overline{\lambda})} \tag{3-28}$$

Barras [115]给出了对 Dnorm 性能的评价，并将 Dnorm 与 Tnorm 结合起来（称为 DTnorm），在 NIST 2003 年的说话人评测上取得了较好的效果。要加入 Dnorm，只需对式（3-26）和（3-27）做如下修改：

$$\Delta S'_{12}=\frac{L(X\mid S_1)-L(X\mid UBM)}{KL(S_1,UBM)}-\frac{L(X\mid S_2)-L(X\mid UBM)}{KL(S_2,UBM)} \tag{3-29}$$

$$\Delta S'_{21}=\frac{L(X\mid S_2)-L(X\mid UBM)}{KL(S_2,UBM)}-\frac{L(X\mid S_1)-L(X\mid UBM)}{KL(S_1,UBM)} \tag{3-30}$$

并将初始聚类和迭代归类算法中所有的ΔS_{12} 和ΔS_{21} 分别替换为 $\Delta S'_{12}$ 和 $\Delta S'_{21}$。

3.2.3.3 聚类实验结果比较分析

在初始聚类时，第 2 步和第 3 步中选取的语音段（即用于生成说话人模型的第一段语音）是否正确对于最后的聚类结果来说是非常重要的，表 3.3 给出了这两步的语音段选取错误率。

从表 3.3 中可以看出，对初始训练模型语音段的选取错误率还是比较低的，因而能够保证后面聚类结果的正确率。

表 3.3 初始聚类的第 2、3 步中语音段的选取错误率

错误类型	错误率（%）
PMissSpkr	0.1
PFASpkr	0.3
PErrSpkr	0.4

表 3.4 给出了初始聚类中使用 Dnorm 和不使用 Dnorm 的聚类错误率比较。

表 3.4 初始聚类中 Dnorm 选用与否的聚类错误率比较

是否使用 Dnorm	错误率（%）
否	11.4
是	10.8

表 3.5 给出了提出的聚类算法(初始聚类和迭代归类)中使用 Dnorm 和不使用 Dnorm 的聚类错误率比较。

表 3.5 聚类中 Dnorm 选用与否的聚类错误率比较

是否使用 Dnorm	错误率（%）
否	6.6
是	5.9

从表 3.4 和 3.5 中可以看出，使用 Dnorm 能够进一步降低算法的聚类错误。这是由于 Dnorm 能够使真实说话人的得分均值与假冒者的得

分均值有较明显的差别，从而降低说话人自身差异对语音段得分的影响。

3.2.4 重分割

从前面的实验结果可以看到，在经过初始分割和聚类之后，仍然存在一定的错误。这些错误中的一部分是由初始分割中产生的漏检错误带来的。由于聚类是在初始分割的结果上进行的，没有对分割点进行增加和减少，因此初始分割产生的漏检错误在聚类时不能得到消除。为了降低漏检错误率，我们在聚类之后，用得到的说话人模型 S_1 和 S_2 对原始多人语音进行重新分割。重分割的步骤如下：(1) 首先对原始语音进行分窗处理，在实验里使用的窗长为 0.8s，窗移为 0.4s；(2) 在分窗后，对每窗的语音计算ΔS_{12}。如果ΔS_{12}大于 0，则该窗属于模型 S_1；否则属于模型 S_2。由于窗内的语音可能含有不同的说话人，因此在计算ΔS_{12}时，没有使用 Dnorm 算法。

虽然国内外研究人员在计算聚类错误率时一般不考虑重叠语音的情况，但是对于一个多说话人识别系统来说，如果不能将这些重叠语音段分离出来，必然会在一定程度上降低系统的性能。由于重叠语音段同时含有两个说话人的信息，一般来说，该语音段在两个说话人模型上的得分差比较小。因此，在重分割第 2 步时，同时计算出每窗语音的$|\Delta S_{12}|$，并按照从大到小的顺序进行排列。在完成重分割后，去除其中$|\Delta S_{12}|$排在后 20%的语音窗。为了叙述方便，将这步处理称为“挑帧处理”。需要说明的是，挑帧处理主要是为了降低第 4 章中多人识别的错误率，见 4.3.3 节。

表 3.6 给出了在表 3.5 基础上使用重分割后的聚类错误率比较，并比较了算法在带挑帧处理和不带挑帧处理下的聚类错误率。

表 3.7 给出文献[43]里 LIA、CLIPS 和 ELISA 在 NIST 2002 的 Switchboard 数据库上测试的结果。其中 LIA+CLIPS 表示将 LIA 的分割结果作为 CLIPS 的输入；CLIPS+LIA 表示将 CLIPS 的分割结果作为 LIA 的输入；Fusion+CLIPS 表示将 CLIPS 和 LIA 分割结果中匹配不上的语音段作为 CLIPS 的输入；Fusion+LIA 表示将 CLIPS 和 LIA 分割结果中匹配不上的语音段作为 LIA 的输入。需要说明的是，Fusion+LIA 是在

NIST 2002 年 Switchboard 数据库说话人分割聚类评测中性能最好的系统。从上面的实验结果可以看出，本文提出的基于 UBM 的说话人分割聚类算法在电话交谈语音上取得了 4.5%的聚类错误率，相对于 Fusion+LIA 来说，聚类错误率相对下降了 21.1%。同时，通过挑帧处理，进一步降低了聚类错误率。

表 3.6 基于 UBM 的说话人分割聚类算法中 Dnorm 和挑帧处理选用与否的聚类错误率比较

算法	错误率（%）
不带 Dnorm 和挑帧处理	5.8
带 Dnorm、但不带挑帧处理	4.5
不带 Dnorm，但带挑帧处理	3.4
带 Dnorm 和挑帧处理	2.6

表 3.7 LIA、CLIPS 和 ELISA 的测试结果（ELISA 包括 4 种组合：LIA+CLIPS、CLIPS+LIA、Fusion+CLIPS 和 Fusion+LIA）

系统	错误率（%）
CLIPS	8.6
LIA	7.4
LIA+CLIPS	7.0
CLIPS+LIA	6.0
Fusion+CLIPS	7.6
Fusion+LIA	5.7

既然本文提出的聚类算法是在初始分割的基础上进行的，那么在重分割的结果上进行再次聚类和重分割，或许能够进一步降低算法的聚类错误率。为了验证这一想法，本文做了相应的实验测试，对聚类部分和重分割部分进行了迭代处理，其中迭代次数设为 2。算法使用了 Dnorm，其聚类错误率为 4.4%，与不迭代的算法（即表 3.6 中的“带 Dnorm、但不带挑帧处理”的算法）相比，错误率基本没有太大的下降，而时间花

销却增加了一倍。因此，在本文的分割聚类算法里，没有对聚类和重分割进行迭代处理。

3.3 小结

本章重点讨论了电话交谈语音中短语音较多的现象，以及它给说话人分割聚类算法带来的影响和可行的解决方案。在此基础上，提出了一种基于 UBM 的说话人分割聚类算法。该算法包括三个步骤：初始分割、聚类和重分割。

在初始分割阶段，考虑到 UBM 能够描述大多数说话人发音分布的特性，采用 UBM 上的对数似然比分来作为电话交谈语音分割的一种度量准则，并利用 BIC 来对分割后的语音段进行合并判决，以降低算法的分割错误率。实验结果表明，基于 UBM 对数似然比分的说话人分割算法，在含有较多短语音段的电话双人交谈语音下，取得了不错的分割效果。

在聚类阶段，使用了说话人模型间的分数差来作为一种将语音段按话者身份进行归类的判断准则。考虑到分割后的语音段平均长度不是很长，因此在初始聚类时，使用了较小混合的 UBM；在初始聚类的基础上，由于可用于训练模型的语音相对较长，因此使用了较大混合的 UBM 来进一步降低聚类错误率。由于说话人自身发音差异的影响，会使该人不同语音段的得分变化较大，从而影响聚类的准确性。为了解决这一问题，在计算模型间分数差时，使用 Dnorm 算法。该算法可以在一定程度上降低说话人自身发音差异对语音段得分带来的影响。

在重分割阶段，首先对原始语音按照一定的窗长和窗移进行分窗处理，然后根据聚类阶段得到的说话人模型，对每窗语音进行归类。重分割可以降低初始分割时产生的漏检错误。为了降低重叠语音对最后说话人识别率的影响，采用了挑帧处理方式，去除了一些重叠语音段。

本章提出的基于 UBM 的说话人分割聚类算法假定电话交谈语音只含两个说话人，而在实际情况下，可能会含有更多的说话人。在预先知道语音中含有的说话人数目时，可以在初始聚类时增加相应的步骤来建立多个说话人的模型；在语音中说话人数目未知的情况下，或者可以采

用聚类的方式估计出语音中的说话人数目,或者可以采用类似 LIA 搜索的方式来估计说话人的数目。对未知说话人数目的聚类研究将是以后的一项研究课题。

第4章 信道鲁棒的说话人识别研究

面向电话信道应用的多说话人识别系统，除了会遇到背景噪音、多说话人问题外，还会遇到信道差异的问题。这一问题会在一定程度上影响说话人识别系统的性能，阻碍说话人识别技术走向实际应用。信道差异是由人们使用的通讯设备（如固定电话、手机）的种类、型号以及通讯途径（如 GSM、CDMA、小灵通等）的不同而产生的，它会使原始语音信号受到不同程度的影响，导致识别系统的性能下降。针对信道差异问题，通常从以下三个方面进行研究：

（1）特征域：常用的方法有倒谱均值减[67]、方差归一化、短时高斯化（Short-Time Gaussianization）、RASTA 滤波、特征弯折[68]和特征映射[69]等；这类算法主要是对特征参数中的信道影响进行消除或补偿。

（2）模型域：常用的方法有说话人模型合成[70]、LFA[72,73]和 NAP[74,75]等。前两个算法是基于 GMM 的。其中，说话人模型合成假定 GMM 的各混合间是相互独立的，通过将 *A* 信道下的模型变为 *B* 信道下的模型来对 *B* 信道下的测试语音进行身份识别。该算法属于模型补偿方面的信道鲁棒算法。而 LFA 也是从信道补偿的角度出发的，但是 LFA 认为 GMM 间的各混合是相关的，因此它将 GMM 各混合的均值向量连接起来构成一个超向量，在该超向量所属的空间上对说话人模型进行信道补偿；NAP 是基于 GMM-SVM 系统的，也是在超向量上进行分析的，但它是从消除模型中干扰说话人识别的信息这一角度出发的。

（3）分数域：常用的有 Hnorm[76]、HTnorm 和 Cnorm[69]等；这类算法主要是通过估计特定信道下的假冒者语音在分数域上的得分分布（通常是单高斯分布），来对该信道下的测试语音的得分做归一化处理，以此减少信道差异对分数值的影响。

虽然上述算法经过实验测试，分别在一定程度的提高了系统的信道鲁棒性，但其中某些算法的融合并不一定能保证提高系统的性能。比如（1）在特征域，短时高斯化、方差归一化和特征弯折常常只选其一；（2）如果在特征域上使用了特征映射，就没有必要在模型域上再使用

说话人模型合成算法，因为两者的作用在理论上是等价的[69]；(3) 分数域上的归一化方法又可以分为基于训练学习的（如 Znorm[83]、Hnorm 和 Dnorm[85]）和基于测试估计的（如 Tnorm[84]）两类，通常两类间的融合（如 DTnorm [115]）能够进一步提高系统的性能。另外，上述算法大多数都需要大量的先验数据，这在一定程度上会给实际应用带来困难。

本章在对 LFA 和 NAP 进行比较分析后，将 LFA 中信道补偿的思想和 NAP 中子空间投影的思想结合起来，提出了一种基于信道子空间投影的模型补偿算法。该算法将测试语音在信道子空间上的投影补偿到说话人模型上，以此减轻信道不匹配所带来的影响。

本章的内容安排如下：第一小节简单介绍一些说话人识别系统中常用的信道鲁棒算法，并给出相应的分析；第二小节介绍提出的基于信道子空间投影的模型补偿算法；第三小节给出实验结果和分析；最后一节是对本章内容的总结。

4.1 常用算法

4.1.1 倒谱均值减

倒谱均值减 (CMS) [67]用来消除信道产生的平稳卷积噪音干扰，公式如下：

$$C_d^{'}(t) = C_d(t) - \frac{1}{N}\sum_{i=1}^{N} C_d(i); \quad d = 1,2,...,D \qquad (4\text{-}1)$$

其中，$C_d(t)$是第 t 帧第 d 维特征分量，D 是特征的维数，N 是特征的总帧数。通常，CMS 是作用于整个语音文件上的。

4.1.2 倒谱方差归一

倒谱方差归一（Cepstral Variance Normalization，CVN）用来消除信道带来的偏移误差，算法定义如下：

$$C_d^{'}(t)=\frac{C_d(t)}{\sigma_d};\quad d=1,2,...,D \tag{4-2}$$

其中 σ_d 是倒谱特征估计得到的标准方差的第 d 维系数。通常，CVN 是作用于整个语音文件上的，但是对于电话信道的语音来说，由于有较多的背景噪音干扰，在整段语音上做 CVN 效果不是很好[116]，因此一般选择在有效语音上进行。另外可以将倒谱均值减和倒谱方差归一合在一起，作用于有效语音段上，用来提高系统的信道鲁棒性，公式如下：

$$C_d^{'}(t)=\frac{C_d(t)-\frac{1}{N}\sum_{i=1}^{N}C_d(i)}{\sigma_d};\quad d=1,...,D \tag{4-3}$$

4.1.3 特征弯折

特征弯折（Feature Warping）[68]的思想是将得到的特征序列通过累积分布函数（Cumulative Distribution Function，CDF）变化为符合标准正态分布的特征序列，来提高特征对不同信道和噪音的鲁棒性[68]。特征弯折假设倒谱特征各维独立，因此可以对各维单独来处理。首先给定滑动窗的窗长 N（即窗内有 N 帧倒谱特征），对一窗内同维的倒谱系数值按照从小到大的顺序进行排序，如果原来处于窗中心位置的倒谱系数值（设为 x）排序后的位置为 r（在 1 和 N 之间），那么对应的 CDF 值 Φ 可以按照下式得到：

$$\Phi=(r-1/2)/N \tag{4-4}$$

那么原中心位置的倒谱系数值 x 在特征弯折后变为 x'，这里 x' 满足：

$$\Phi=\int_{-\infty}^{x'}f(z)dz \tag{4-5}$$

其中 $f(z)$是标准正态分布的概率密度函数，定义如下：

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \tag{4-6}$$

特征弯折是近些年在说话人识别中提出来的特征归一化方法，在单人跨信道识别中取得了较好的效果。它与 CMS 相结合，能够进一步降低系统的等错误率。

倒谱均值减、方差归一化、短时高斯化（Short-Time Gaussianization）、RASTA 滤波和特征弯折这些算法不需要额外的先验数据，因此被说话人识别系统广泛采用。在本文的实验中，采用了如式（4-3）的特征归一化算法。

4.1.4 说话人模型合成和特征映射

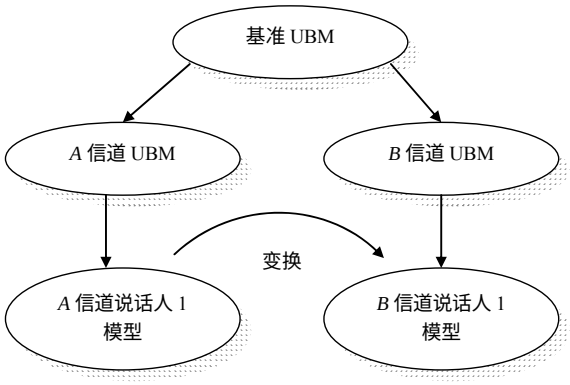

图 4.1 说话人模型合成示意图[69]

说话人模型合成[70]的基本思想是通过将 *A* 信道的说话人模型变换为 *B* 信道的说话人模型，对 *B* 信道的测试语音进行识别，如图 4.1 所示。首先，用大量信道先验数据训练得到基准 UBM，然后用信道相关（如信道 *A* 和 *B*）的先验数据从基准 UBM 上得到 *A* 信道的 UBM 和 *B* 信道的 UBM。然后我们假设已有信道 *A* 下说话人 1 的模型，并且测试语音属于 *B* 信道（可以通过测试语音在 *A* 信道的 UBM 和 *B* 信道的 UBM 上的得分大小进行判断），这时若用 *A* 信道下说话人 1 的模型来识别 *B* 信道下测试语音的话者身份，必然会受到信道差异的影响。那么如何得到说话人 1 在 *B* 信道下的模型呢？我们可以利用已有信道 *A* 和 *B* 的 UBM 间的关系来得到说话人 1 在 *B* 信道下的模型，如下式所示：

$$w_i^{SB} = w_i^{SA} \cdot \left(\frac{w_i^{UB}}{w_i^{UA}} \right) \tag{4-7}$$

$$\mu_i^{SB} = \mu_i^{SA} + \left(\mu_i^{UB} - \mu_i^{UA} \right) \tag{4-8}$$

$$\sigma_i^{SB} = \sigma_i^{SA} \cdot \left(\frac{\sigma_i^{UB}}{\sigma_i^{UA}} \right) \tag{4-9}$$

其中(w_i, μ_i, σ_i)表示第 i 个高斯混合的参数（权重、均值、方差），*SA*、*SB*、*UA* 和 *UB* 分别对应说话人在 *A* 信道下的模型、说话人在 *B* 信道下的模型、*A* 信道 UBM 和 *B* 信道 UBM。

特征映射[69]的基本思想是将不同信道下的特征映射到一个信道无关的特征空间上。跟说话人模型合成类似，首先得到基准 UBM、*A* 信道 UBM 和 *B* 信道 UBM；然后用测试语音在 A 信道 UBM 和 *B* 信道 UBM 中选出可能的信道类型，假设为 *B*；接着求得当前特征帧在 *B* 信道 UBM 上得分最大的混合序号，设为 i，那么我们可以通过下式来将当前特征帧映射到一个信道无关的特征空间上：

$$y = \left(x - \mu_i^{UB} \right) \frac{\sigma_i^{root}}{\sigma_i^{UB}} + \mu_i^{root} \tag{4-10}$$

其中 x 和 y 分别是原始特征和变换后的特征，*root* 表示基准 UBM。

说话人模型合成与特征映射对测试信道未知的情况是无能为力的，但是如果测试信道可知且能够得到大量先验信道数据，这两种算法都能够取得不错的性能。另外，这两种算法是否有效的一个关键是能否正确的检测出测试语音的信道类型，即能否达到一定的信道分类正确率。如果信道分类的效果比较差，说明选取的信道相关数据覆盖面或者不够，需要更多的数据；或者太宽，需要进一步细分信道类型。相对来说，特征映射的应用灵活性要高于说话人模型合成，这是因为：(1) 不需要为每个说话人在不同信道下建立模型；(2) 能够使用多段不同信道下同一说话人的语音来训练一个说话人模型；(3) 能够将不同信道下的语音所

提取出的特征变换到同一信道下，方便后续模块的处理。

在 NIST 2006 年的说话人识别数据库中，由于测试数据含有的信道类型比较多（见 4.3.1 节），需要的先验信道数据量非常大，而本文的实验缺乏足够的先验信道数据（尤其是某些信道的数据非常缺乏），导致训练得到的信道相关 UBM 对测试信道的分类性能较差，因此在实验中没有使用特征映射和说话人模型合成。

4.1.5 LFA 和 NAP

LFA[72,73]和 NAP[74,75]是最近几年提出的信道鲁棒算法，LFA 是从信道补偿的角度出发的，而 NAP 则是从消除模型中干扰说话人识别的信息这一角度出发的。LFA 和 NAP 都是在超向量（Supervector）空间进行分析的。所谓超向量是指将 GMM 中的每个混合均值依次连接起来所构成的向量，其定义如下：

给定一个说话人模型（GMM），它含有 K 个高斯混合，每个混合均值的维数为 F（即特征的维数）。将这 K 个高斯混合的均值按照高斯混合的序号连接起来，组成一个超向量 M，表示如下：

$$M = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mathrm{M} \\ \mu_K \end{pmatrix} \tag{4-11}$$

这里 M 的维数为 $KF \times 1$。在 LFA 和 NAP 中，构成超向量的说话人模型都是从 UBM 上用 MAP 算法得到的。

LFA 算法是由 Patrick 提出的，应用于 GMM-UBM 系统中。Patrick 将超向量 M 定义为：

$$M = S + C \tag{4-12}$$

其中 S 被称为说话人超向量（Speaker Supervector），C 被称为信道超向量（Channel Supervector），两者均服从正态分布，如图 4.2 所示。

式（4-12）中的 S 和 C 分别定义为[117]：

$$S = m + Vy \tag{4-13}$$

$$C = Uz \tag{4-14}$$

这里 m 表示说话人的共性向量,可由 UBM 中的混合均值连接组成，V 和 U 均为低秩矩阵，y 和 z 分别表示说话人因子（Speaker Factor）向量和信道因子（Channel Factor）向量。m 为 $KF\times 1$ 维的超向量，V 和 U 分别为 $KF\times R_S$ 维和 $KF\times R_C$ 维的低秩矩阵，R_S 为 V 矩阵的秩，R_C 为 U 矩阵的秩，y 和 z 分别为 $R_S\times 1$ 维和 $R_C\times 1$ 维的向量。

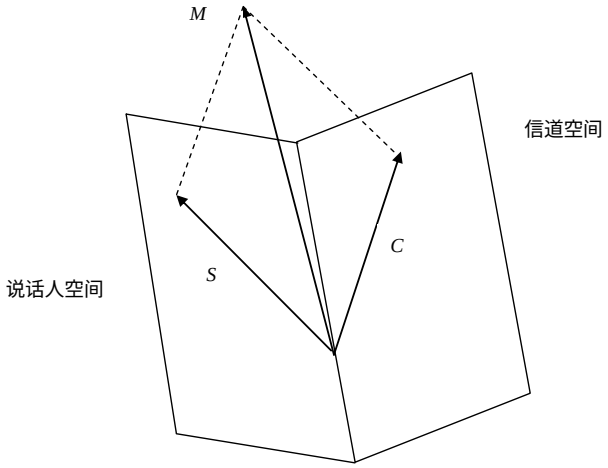

图 4.2 一个说话人和信道相关的超向量 M 可以分解为两个超向量之和，一个在说话人空间（即 S），另一个在信道空间（即 C）[117]

LFA 的基本思想是要估计出语音中含有的特定说话人超向量和信道超向量。在训练时，从 M 中去除 Uz；在测试时，用测试语音估计出的 Uz 来补偿 M，并用补偿后的模型对测试语音进行识别。注意，V 和 U 矩阵是作为两个相互关联的量，从大量集外说话人的语音数据里用 EM 算法估计得到的，可以参看文献[73]。V 和 U 是预先估计得到的，在训练和测试时保持不变。y 和 z 也是作为两个相互关联的量，从训练语音和测试语音中用迭代算法估计得到的，可以参看文献[117]。LFA 不仅仅要计算 V、U、y 和 z，还要计算模型补偿后的混合方差，因此它的计算量非常大，实现起来比较困难。

NAP 算法是由 MIT 林肯实验室的 Campbell 等人提出的，应用于 GMM-SVM 系统中，它的基本思想（如图 4.3 所示）是用 SVM 来分析去除了干扰说话人识别的向量后的超向量 M（即图 4.3 中的 Pm）。图 4.3

中的 m 为如式（4-11）所示的超向量，P 矩阵为投影矩阵，它描述了由干扰说话人识别的向量构成的投影空间，I 矩阵为单位矩阵。P 矩阵是用 PCA 方法从大量集外说话人数据上估计得到的，可以参考文献[74]。一般来说，在 SVM 中选取的核函数为

$$K\left((Pm)^a,(Pm)^b\right)=\sum_{i=1}^{N}\left(\sqrt{w_i}\Sigma_i^{-(1/2)}(Pm)_i^a\right)^t\left(\sqrt{w_i}\Sigma_i^{-(1/2)}(Pm)_i^b\right) \quad (4\text{-}15)$$

其中 a 和 b 表示两段语音，$(Pm)^a$ 为用 a 语音生成的超向量在干扰说话人识别的空间上的投影，w_i，Σ_i 分别为第 i 个高斯混合的权重和协方差矩阵，N 为高斯混合的个数，$(Pm)_i^a$ 为 $(Pm)^a$ 中第 i 个高斯混合的均值向量，$(\cdot)^t$ 表示转置运算。

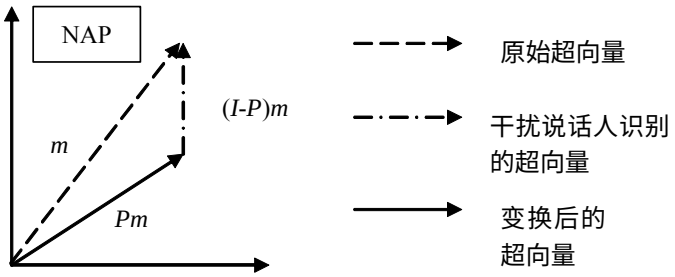

图 4.3 NAP 算法基本思路示意图[118]

LFA 和 NAP 在 NIST 2006 年说话人识别评测中取得了很好的效果，有关这两种算法更详细的说明可以参看文献[72~75]。

4.1.6 Hnorm、Tnorm 和 Znorm

Znorm[83]在上世纪九十年代中期被广泛使用，它的基本步骤是：(1) 在训练时，用一个真实说话人模型和一组假冒者语音来估计假冒者在该说话人模型上的得分分布，即求出得分的均值和标准方差；(2) 在测试时，用得到的均值和标准方差对该说话人模型的得分按下式进行归一化：

$$L'(X\mid\lambda)=\frac{L(X\mid\lambda)-\mu}{\sigma} \quad (4\text{-}16)$$

其中 X 是测试语音，λ 是说话人模型，μ和σ是估计得到的均值和标准方差，$L(X|\lambda)$为 X 在模型 λ 上的似然分。

在电话语音测试中，人们发现，虽然传输信道一致，但是由于听筒类型的不同，也会较大的降低系统的识别性能。为此，Reynolds 提出了一种 Znorm 的变形：Hnorm[76]。在训练时，用不同听筒类型的假冒者语音集来估计出说话人模型在不同听筒类型下的得分分布；在测试时，根据测试语音所处的听筒类型，选用相应的归一化参数，按照式（4-16）来对测试语音的得分进行归一化处理。

Tnorm[84]是目前被广泛使用的一种分数归一化方法。首先给定一组假冒者的模型，在测试的时候，用测试语音在这些假冒者模型上的得分估计出均值和标准方差，然后按照式（4-16）来对说话人模型的得分进行归一化处理。如果用跟测试语音信道一致的假冒者模型集来估计假冒者分数的均值和标准方差，那么这种 Tnorm 被称为 HTnorm。Sturim 等[119]在 2005 年提出了一种 Tnorm 的改进算法 ATnorm，该算法在训练时，用 N 段假冒者语音从一个含有较多假冒者模型的集外模型池 P 中找出与当前目标说话人模型 S 得分最接近的 K 个假冒者模型，在测试时，使用这 K 个模型的得分分布来对测试语音在模型 S 上的得分进行归一化处理。ATnorm 的效果要比 Tnorm 更好，但是它的计算复杂度和对数据量的要求要比 Tnorm 高很多。这里需要说明的是，Znorm、Tnorm、Dnorm 不属于专门针对信道问题的分数归一化算法。

上述几种分数归一化算法要求的数据量都比较大，在实际应用中有一定困难，在本文的实验里，使用了 Tnorm 算法。

4.1.7 评测标准

评价一个说话人识别系统的识别性能可以看它的两个错误率：错误接受率（False Acception Rate，FAR；也被称为 False Alarm Rate）和错误拒绝率（False Rejection Rate，FRR；也被称为 Miss Probability）。前者指的是一个系统对集外说话人的接受性能，该值越低，说明系统越安全，不易被闯入；后者指的是一个系统对真实说话人的拒绝性能，该值越低，说明真实说话者越容易进入系统。这两个错误率跟判决阈值有关，阈值越低，系统的 FRR 越低，相应的 FAR 就越高；阈值越高，系统的

FRR 越高，相应的 FAR 就越低。也就是说，FAR 和 FRR 都是判决阈值的函数，这两个函数在值域相交的点称为等错误率点（Equal Error Rate Point)。通常人们希望系统的等错误率尽可能低，也就是 FAR 和 FRR 相等时的值尽可能小。研究人员常常使用检测错误权衡曲线（Detection Error Trade-offs Curve，DET Curve）[120]来反映这两个错误率之间的关系。在 DET 曲线上，曲线越接近原点，系统的识别性能越好。

除了可以使用 DET 曲线和等错误率来反映系统的识别性能好坏，NIST 还定义了 FAR 和 FRR 的加权和函数，该函数被称之为检测代价函数（Detection Cost Function，DCF)。通常，针对不同的应用背景，对 FAR 和 FRR 定义不同的权重（代价），并用最小 DCF 来表示系统能够取得的最优性能。因此，对于面向实际应用的系统来说，最小 DCF 要比 EER 更有意义。在给定不同错误率权重（代价）下，最小 DCF 越小，系统的实际应用性能越好。DCF 的定义如下：

$$C_{Det} = C_{Miss} \times P_{Miss} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm} \times \left(1 - P_{Target}\right) \qquad (4\text{-}17)$$

其中，C_X 表示不同错误率的权重，P_X 表示不同的错误率，P_{Target} 表示目标说话人的先验概率。在 NIST2002 年的评测中，C_{Miss}、$C_{FalseAlarm}$ 和 P_{Target} 的取值分别为 10、1 和 0.01，后面实验计算最小 DCF 时，如不特别说明，也使用同样的数值。

4.2 基于信道子空间投影的模型补偿算法

在 4.1 节里介绍的 LFA 和 NAP 算法认为 GMM 中的高斯混合是相关的，并在如式（4-11）所示的超向量组成的空间上分析信道对模型的影响。但这两种方法的出发点是不同的，LFA 是对信道进行补偿，而 NAP 则是消除模型中干扰说话人识别的信息。相对来说，NAP 要比 LFA 更容易实现，但需要其他的分类器（如 SVM)，而 LFA 的效果要稍好于 NAP，可以应用于 GMM-UBM 系统中。本文在对 LFA 和 NAP 进行分析后，将 LFA 中信道补偿的思想和 NAP 中子空间投影的思想结合起来，提出了基于信道子空间投影的模型补偿算法。该算法的基本思想是采用投影的方式来估计测试语音在模型空间中含有的与信道相关的信息，并

用该信息对训练得到的说话人模型进行补偿。这样，算法一方面避免了 LFA 中说话人因子和信道因子的复杂计算，实现起来比较简单，另一方面又能够应用于 GMM-UBM 系统，不需要额外的分类器（如 NAP 中的 SVM）。

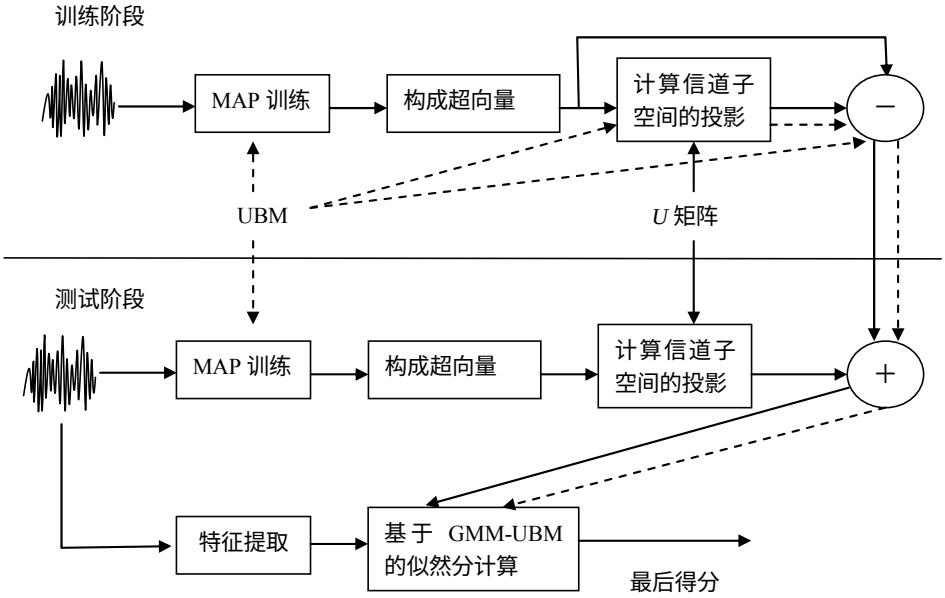

图 4.4 基于信道子空间投影的模型补偿算法在 GMM-UBM 系统上的应用

本文所提算法在 GMM-UBM 系统中的应用流程如图 4.4 所示。图 4.4 中的 U 矩阵为信道子空间的投影矩阵。基于信道子空间投影的模型补偿算法可以分为四步：(1) U 矩阵的估计：用 PCA 算法从大量集外说话人数据中估计出信道子空间的一组正交基，构成矩阵 U；(2) 训练说话人模型：首先用训练语音从 UBM 上采用传统的 MAP 算法得到说话人模型，并构成超向量，接着计算出该向量在信道子空间的投影，最后保留去除了该投影的说话人模型；(3) 补偿说话人模型：首先用测试语音从 UBM 上采用传统的 MAP 算法得到说话人模型，并构成超向量，接着计算出该向量在信道子空间的投影，最后用该投影来补偿训练得到的说话人模型；(4) 识别测试语音：用补偿后的说话人模型按照传统的 GMM-UBM 识别方式来判断测试语音的话者身份。注意：在用 MAP 算

法生成模型时，使用的是原始的 UBM；在识别测试语音时，使用的是补偿后的 UBM。4.2.1 节至 4.2.4 节将详细介绍这四步。

4.2.1 U 矩阵的估计

给定一个说话人 s 的第 i 段语音，它所处的信道类型为 c_i，用传统的 MAP 算法从 UBM 上自适应得到 s 的模型 $M(s,i,c_i)$。$M(s,i,c_i)$为形如式（4-11）的超向量，可以表示如下：

$$M(s,i,c_i)=m(s)+Uz(s,c_i) \tag{4-18}$$

其中 $m(s)$描述的是与说话人 s 相关的超向量，而 U 矩阵为一低秩矩阵，是由信道子空间的一组正交基构成的，$z(s,c_i)$为信道因子。这里，$M(s,i,c_i)$和 $m(s)$为 $KF\times 1$ 的超向量，U 为 $KF\times R_C$ 的矩阵，R_C 为 U 矩阵的秩，$z(s,c_i)$为 $R_C\times 1$ 的向量，$Uz(s,c_i)$描述了说话人 s 第 i 段语音在信道子空间上的投影。由于声学特征经归一化处理后，服从标准正态分布，因此 $z(s,c_i)$也应该服从标准正态分布。给定说话人 s 的 N 段语音（所处的信道类型可能相同，也能不同），在 N 较大的情况下，式（4-18）中的 $m(s)$可以按照下式求得：

$$\bar{M}(s)=\frac{1}{N}\sum_{i=1}^{N}M(s,i,c_i)\approx m(s) \tag{4-19}$$

这样，说话人 s 的第 i 段话在信道子空间上的投影 $Uz(s,c_i)$可按下式求得：

$$Uz(s,c_i)=M(s,i,c_i)-\bar{M}(s) \tag{4-20}$$

注意，即使说话人 s 的第 i 段话和第 j 段话所处的信道类型相同，$z(s,c_i)$和 $z(s,c_j)$也是不完全相同的。

由上述投影向量构成的自相关矩阵 A_s 可以表示如下：

$$A_s=\sum_{i=1}^{N}Uz(s,c_i)z(s,c_i)^tU^t \tag{4-21}$$

其中 A_s 为 $KF\times KF$ 的矩阵。虽然 U 矩阵可以用 PCA 方法从 A_s 中求得，但是这样求得的 U 矩阵不能精确的描述出信道子空间。

为了求得更加精确的 U 矩阵，可以对由多个说话人的 A_s 矩阵所构成的平均类内自相关矩阵 A 进行 PCA 分析，用求出的最大前 R_C 个特征值所对应的特征向量来构成 U 矩阵。给定 L 个说话人，每个说话人有 N 段话的情况下，A 矩阵可以表示为：

$$A=\frac{1}{L}\sum_{i=1}^{L}A_{s_i}=\frac{1}{L}\sum_{i=1}^{L}\sum_{j=1}^{N}Uz\left(s_i,c_j\right)z\left(s_i,c_j\right)^tU^t \tag{4-22}$$

这里 A 为 $KF\times KF$ 的矩阵。对 A 进行 PCA 分析，用求得的最大前 R_C 个特征值所对应的特征向量来构成 U 矩阵，即：

$$U=\begin{pmatrix}v_1 & v_2 & \mathrm{L} & v_{R_C}\end{pmatrix} \tag{4-23}$$

其中 v_i 表示第 i 大的特征值所对应的特征向量，其维数为 $KF\times 1$，U 矩阵的维数为 $KF\times R_C$。

注意，这里组成 U 矩阵的特征向量是经过正交化处理过的，即 $U^tU=I$，I 为 $R_C\times R_C$ 的单位矩阵。那么，由 U 矩阵构成的投影矩阵 P 可以表示为：

$$P=UU^t，\text{且 } PU=UU^tU=U \tag{4-24}$$

这里 P 矩阵的维数为 $KF\times KF$，它描述了信道子空间。

4.2.2 训练说话人模型

在训练时，给定说话人 s 的一段训练语音 i，所处的信道类型为 c_i。首先用传统的 MAP 算法从原始 UBM 上自适应得到说话人 s 的模型，然后构造超向量 $M(s,i,c_i)$，并计算该超向量在信道子空间的投影，即：

$$C\left(s,c_i\right)=PM\left(s,i,c_i\right)=Pm\left(s\right)+Uz\left(s,c_i\right) \tag{4-25}$$

最后，从 $M(s,i,c_i)$中去除 $C(s,c_i)$后，作为说话人 s 的模型保留下来，即：

$$\begin{aligned} M'(s) &= (I-P)M(s,i,c_i) \\ &= (I-P)m(s) \end{aligned} \tag{4-26}$$

这里 I 为 $KF\times KF$ 的单位矩阵，描述全空间。

对于原始 UBM 来说，其混合均值构成的超向量可以表示为：

$$M(ubm) = m \tag{4-27}$$

接着去除该超向量在信道子空间上的投影：

$$M'(ubm) = (I-P)m \tag{4-28}$$

最后，将 $M'(ubm)$ 保留下来。

4.2.3 补偿说话人模型

在测试时，给定说话人 t 的一段测试语音 j，所处的信道类型为 c_j。首先用传统的 MAP 算法从原始 UBM 上自适应得到说话人 t 的模型，然后构造超向量 $M(t,j,c_j)$，并计算该超向量在信道子空间上的投影，即

$$C(t,c_j) = PM(t,j,c_j) = Pm(t) + Uz(t,c_j) \tag{4-29}$$

为了得到说话人 s 在信道 c_j 下的模型，需要对 $M'(s)$ 进行补偿，即在 $M'(s)$ 上加上 $C(t,c_j)$，即

$$M(s,c_j) = M'(s) + C(t,c_j) \tag{4-30}$$

将式（4-26）和式（4-29）带入式（4-30）中可以得到：

$$M(s,c_j) = (I-P)m(s) + Pm(t) + Uz(t,c_j) \tag{4-31}$$

这里，$M(s,c_j)$ 即为说话人 s 在测试语音 j 所处信道 c_j 下的模型。

为了得到在测试语音 j 所处信道 c_j 下的 UBM，可以用 $C(t,c_j)$ 对 $M'(ubm)$ 进行补偿，即

$$\begin{aligned} M\left(ubm,c_j\right) &= M'\left(ubm\right)+C\left(t,c_j\right) \\ &= \left(I-P\right)m+Pm\left(t\right)+Uz\left(t,c_j\right) \end{aligned} \tag{4-32}$$

这样就得到了在测试语音 j 所处信道 c_j 下的特定 UBM 和说话人模型，下面就可以按照传统的 GMM-UBM 识别系统对测试语音进行身份判断。

4.2.4 识别测试语音

首先，我们来看看信道差异对未进行信道补偿处理的说话人识别系统的影响。

给定一段测试语音 j，所处的信道类型为 c_j，设该语音所提取的特征序列为 $\{f_t, t=1,2,\cdots,T\}$，其中第 t 帧特征与该段语音所生成的模型的第 k 个混合最接近，则 f_t 可以表示为：

$$f_t = \left[m\left(t\right)\right]_k + \left[Uz\left(t,c_j\right)\right]_k + \Sigma_k d \tag{4-33}$$

其中 $[.]_k$ 表示取超向量中序号为 $(k\text{-}1)F$ 到 $kF\text{-}1$ 的元素组成的维数为 $F\times 1$ 的向量，F 为特征的维数，Σ_k 为说话人模型中第 k 个高斯混合的协方差矩阵，d 为一个服从标准正态分布的随机变量。给定由说话人 s 的原始 GMM 模型构成的超向量 $M(s,i,c_i)$，那么 f_t 在第 k 个高斯混合上的得分为：

$$\begin{aligned} H\left(f_t \mid \left[M\left(s,i,c_i\right)\right]_k\right) &= \frac{1}{\left(2\pi\right)^{F/2}\left|\Sigma_k\right|^{1/2}} \\ &\cdot \exp\left\{-\frac{1}{2}\left(f_t-\left[M\left(s,i,c_i\right)\right]_k\right)^t \Sigma_k^{-1}\left(f_t-\left[M\left(s,i,c_i\right)\right]_k\right)\right\} \end{aligned} \tag{4-34}$$

式（4-34）主要取决于下式的大小

$$\left(f_t-\left[M\left(s,i,c_i\right)\right]_k\right)^t \Sigma_k^{-1}\left(f_t-\left[M\left(s,i,c_i\right)\right]_k\right) \tag{4-35}$$

将式（4-33）带入式（4-35）得

$$\begin{aligned}&\left(\left[m(t)\right]_k+\left[Uz\left(t,c_j\right)\right]_k+\Sigma_k d-\left[m(s)\right]_k-\left[Uz\left(s,c_i\right)\right]_k\right)^t\\&\cdot\Sigma_k^{-1}\cdot\left(\left[m(t)\right]_k+\left[Uz\left(t,c_j\right)\right]_k+\Sigma_k d-\left[m(s)\right]_k-\left[Uz\left(s,c_i\right)\right]_k\right)\end{aligned}\tag{4-36}$$

若测试语音与说话人 s 的模型所处的信道类型一致，即 $Uz(t,c_j)\approx Uz(s,c_i)$，则上式可以近似表示为：

$$\left(\left[m(t)\right]_k+\Sigma_k d-\left[m(s)\right]_k\right)^t\Sigma_k^{-1}\left(\left[m(t)\right]_k+\Sigma_k d-\left[m(s)\right]_k\right)\tag{4-37}$$

从式（4-37）可以看出，如果测试语音和训练语音的信道相匹配，测试语音在说话人模型上的得分能够较好的反映测试语音和说话人模型的匹配程度。

如果测试语音与说话人 s 的模型所处的信道类型不匹配，则测试语音在说话人模型上的得分会受到 $Uz(t,c_j)$和 $Uz(s,c_i)$的影响（如式（4-36）所示），不能很好的反映它们之间的匹配程度。

下面来看一下用本文提出的算法对说话人模型进行信道补偿后，再对测试语音进行识别的情况。

将式（4-31）中的 $M(s,c_j)$替换式（4-35）中的 $M(s,i,c_i)$后得：

$$\left(f_t-\left[M\left(s,c_j\right)\right]_k\right)^t\Sigma_k^{-1}\left(f_t-\left[M\left(s,c_j\right)\right]_k\right)\tag{4-38}$$

化简后得：

$$\begin{aligned}&\left(\left[(I-P)\left(m(t)-m(s)\right)\right]_k+\Sigma_k d\right)^t\\&\cdot\Sigma_k^{-1}\left(\left[(I-P)\left(m(t)-m(s)\right)\right]_k+\Sigma_k d\right)\end{aligned}\tag{4-39}$$

从式（4-39）可以看到，测试语音中含有得 $Uz(t,c_j)$得到了消除，从而减少了信道差异对测试语音得分的影响。虽然式(4-39)中引入了$(I-P)$这一项，但由于它在测试语音对每个说话人模型的算分过程中保持不变，因此该项对于测试语音得分的影响不是很大。从式（4-39）可以看到，本文提出的基于信道子空间投影的模型补偿算法能够在一定程度上消除信道差异对识别性能的影响。

4.3 实验结果与分析

为了评测本章提出的信道鲁棒算法以及前几章提出的算法在电话信道下的识别性能，本章实验使用了 NIST 2006 年的说话人评测数据库，从中选取了两个评测任务所用到的数据库：单人训练单人识别数据库（对应于 1conv4w-1conv4w）和单人训练双人识别数据库（对应于 1conv4w-1conv2w）[121]。这两个数据库的语音文件均为 Sphere 格式，其中的语音数据格式为 8kHz 采样率，8 位精度，mu-law 压缩。下面一节将简单介绍一下这两个数据库的组成情况，为了方便后面实验结果的说明，将这两个识别任务分别简称为单人识别和双人识别。

4.3.1 实验设置和数据库

这两个数据库都是在实际电话信道下录制的，从传输信道上来看，可以大致分为三类：手机（Cellular）、无绳电话（Cordless）和座机（Landline）。每个信道根据语音采集设备的不同，又可以分为四个小类：Speaker-Phone、Head-mounted、Ear-bud 和 Hand-held。但是上面的信道分类并不全面，例如手机信道又可以细分为 CDMA 和 GSM 等。因此，说话人识别系统如果要加入特征映射或是说话人模型合成这些算法，所需要的先验信道相关的数据量是非常大的。

对于单人测试来说，训练语音和测试语音均为双声道，其中目标说话人共有 816 人（女声 462、男声 354）。测试语音共 2467 段，总共有 53966 次确认测试（即一段测试语音与一个模型进行身份确认）。

对于双人测试来说，训练语音与单人库一样，测试语音为单声道，总共有 52883 次确认测试。

需要说明的是，每段语音的话者性别信息是已知的，语音长度为 5 分钟（包含静音部分），而且上面的确认测试都是同性别的测试，即男声语音测试男声模型，女声语音测试女声模型。另外，在这两个数据库中，同一个说话人的不同语音所属的语种可能不同，如英语、俄语、汉语等。

实验中 UBM 的训练数据来自 NIST 2004 年的测试数据库，系统共使用了两个性别相关的 UBM（即男、女 UBM），每个 UBM 有 1024 混合，分别用 8 小时的语音（含静音部分）训练得到。系统所用的 VAD

方法是基于能量的。用于 Tnorm 的语音为男声 245 段、女声 368 段（每个说话人一段话），它们来自 NIST 2005 年的数据库。

在本章的实验里，用于估计 U 矩阵的数据来自 NIST 2005 年的说话人识别数据库，其中有 295 个女性说话人和 202 个男性说话人，每个说话人 8 段话，每段话 5 分钟（含静音部分）。按照式（4-22）求得两个男、女性别相关的 A 矩阵，图 4.5 给出了男、女的 A 矩阵经过 PCA 分析后特征值大小的变化曲线，图中只给出了最大的前 300 个特征值。从图 4.5 中可以看到，随着序号的增加，特征值的大小下降很快，在序号 50 之后，基本变化不是很明显。因此在实验里，男、女相关的 U 矩阵的秩取为 50，即 R_C 为 50。为了验证 R_C 的选择是否合适，本章进行了专门的实验，可以参看 4.3.2 节的图 4.8 和表 4.3。

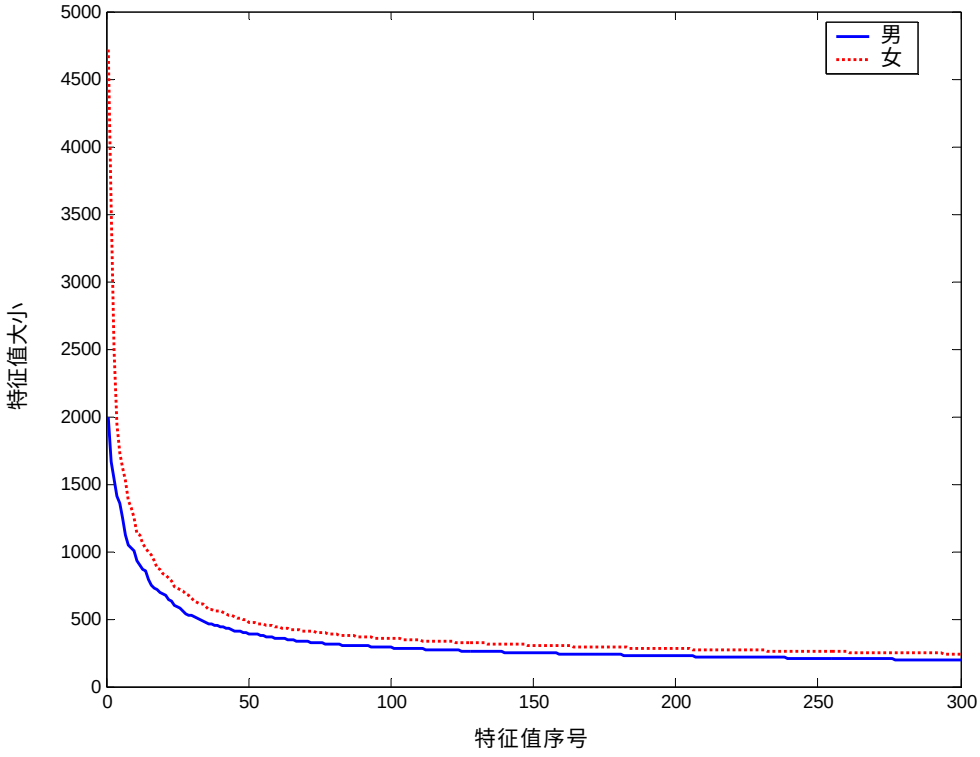

图 4.5 男、女 R 矩阵对应的最大前 300 个特征值的大小变化曲线图

4.3.2 实验一 单人识别

图 4.6 给出了不带任何信道鲁棒算法的基准系统（Baseline）、Baseline＋Tnorm、基于信道子空间投影的模型补偿算法的系统（简称为 CSPBMC）和 CSPBMC＋Tnorm 这四个系统在单人识别库上的性能比较。这里，CSPBMC 系统中的 R_C 为 50。表 4.1 给出了上述系统的另外两个评测指标：最小 DCF 和 EER。

从图 4.6 和表 4.1 可以看出，CSPBMC 算法能够有效的降低说话人识别系统的识别错误率。相对于 Baseline 系统来说，等错误率相对下降了 15.4%，相对于 Baseline+Tnorm 来说，等错误率相对下降了 6.3%。CSPBMC 与 Tnorm 相结合，进一步降低了说话人识别系统的等错误率，相对于 Baseline+Tnorm 来说，等错误率相对下降了 16.2%，DCF 相对下降了 23.4%。

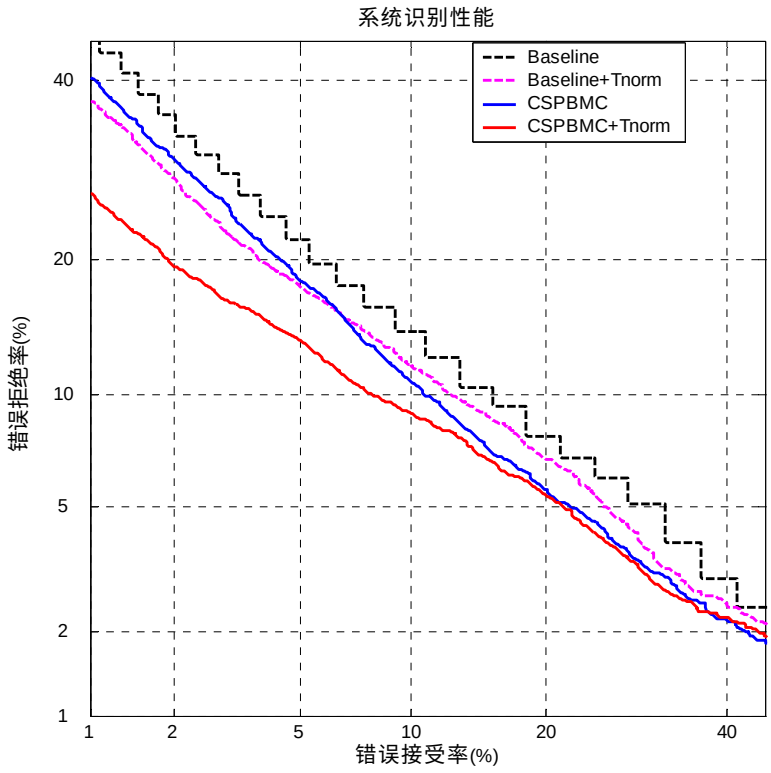

图 4.6 四种系统在单人识别实验中的 DET 曲线对比

表 4.1 四种系统在单人识别实验中的系统性能对比

系统	最小 DCF($\times 10^{-2}$)	EER(%)
Baseline	5.3	12.3
Baseline+Tnorm	4.7	11.1
CSPBMC	4.8	10.4
CSPBMC+Tnorm	3.6	9.3

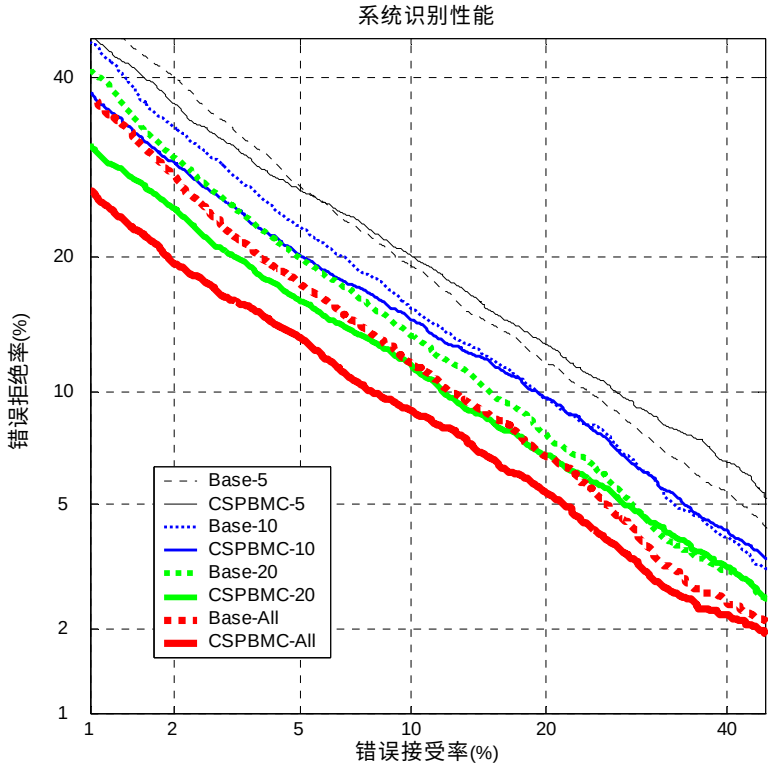

图 4.7 两种系统在 4 组测试语音长度下的 DET 曲线对比

由于 CSPBMC 算法是在说话人模型空间上估计训练语音或测试语音在信道子空间上的投影，因此说话人模型的精度会在很大程度上影响算法的性能，而说话人模型的精度又往往取决于训练语音的有效长度。为了评价 CSPBMC 算法在不同长度的测试语音下的性能，我们安排了四组实验（即测试语音的有效长度为 5 秒、10 秒、20 秒和所有有效语

音），并比较了两个系统。这两个系统分别对应于图 4.6 中的 Baseline+Tnorm（记为 Base）和 CSPBMC+Tnorm（记为 CSPBMC）。图 4.7 和表 4.2 给出了这两个系统在单人识别数据库上的性能比较。

表 4.2 两种系统在 4 组测试语音长度下的系统性能对比

系统	最小 DCF($\times 10^{-2}$)	EER(%)
Base-5/ CSPBMC -5	5.8/5.5	14.8/15.5
Base-10/ CSPBMC -10	5.2/4.7	13.2/12.9
Base-20/ CSPBMC -20	4.9/4.1	12.1/10.9
Base-All/ CSPBMC -ALL	4.7/3.6	11.1/9.3

从实验结果中，我们可以看到：(1) 在测试语音有效长度为 5 秒的情况下，用测试语音估计得到的说话人模型精度不高，导致该测试语音在信道子空间上的投影估计不够准确，从而使得系统的识别性能有所下降；(2) 在测试语音有效长度为 10 秒的情况下，用测试语音估计得到的说话人模型精度仍有不足，因此 CSPBMC 的性能与 Base 系统的等错误率相差不多，但最小 DCF 有较大的降低；(3) 在测试语音有效长度为 20 秒的情况下，用测试语音估计得到的说话人模型精度已经能够有所保证，因此 CSPBMC 的系统性能要明显好于 Base 系统，这也说明在使用传统 MAP 算法的时候，要想得到较精确的模型，测试语音的有效长度不能太短。

为了验证本章实验里使用的基于信道子空间投影的模型补偿算法所选取的 R_C 值（即 U 矩阵的秩，或信道子空间的维数）是否最优，本文进行了相应的实验。实验中使用的系统是基于 CSPBMC＋Tnorm 的，测试了该系统在 R_C 为 10、30、50 和 100 这四个数值下的系统性能，实验结果可以见图 4.8 和表 4.3。从实验结果中可以看到，系统最优性能下的 R_C 值与从图 4.5 中特征值的变化曲线中得到的结果是一致的。从图 4.8 和表 4.3 可以看到，当 R_C 的值过小时，会使得 U 矩阵不够精确；当 R_C 的值过大时，会使得估计出的信道子空间投影含有较多特定说话人相关的信息，影响系统的识别性能。

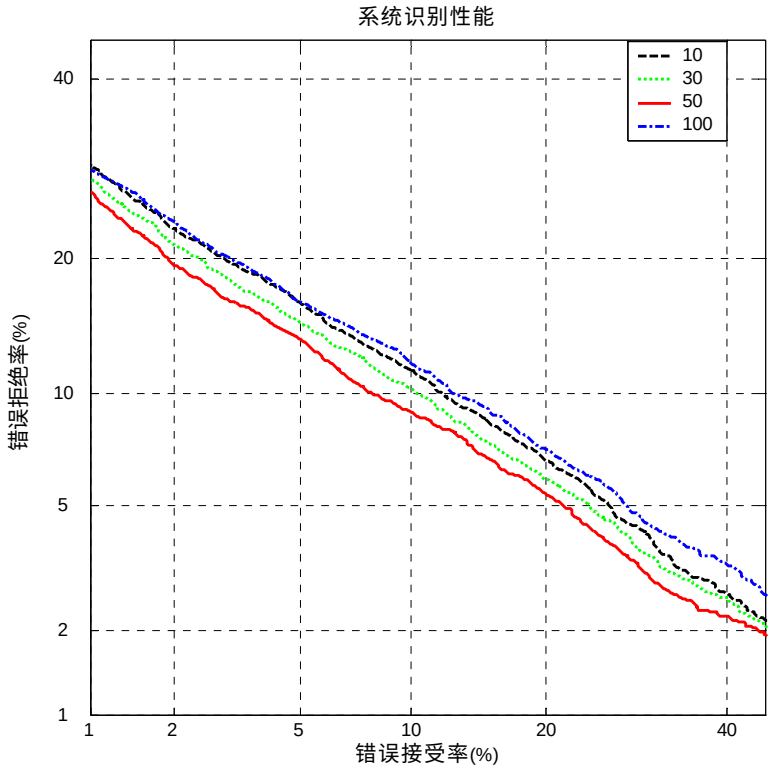

图 4.8 不同 R_C 数值下的系统 DET 曲线对比

表 4.3 不同 R_C 数值下，CSPBMC＋Tnorm 系统的性能对比

R_C	最小 DCF($\times 10^{-2}$)	EER(%)
10	3.9	10.8
30	3.7	10.1
50	3.6	9.3
100	3.8	11.2

4.3.3 实验二 双人识别

为了测试论文所提出的三个算法在电话信道下双说话人识别上的整体性能，构建了一个完整的双人识别系统，如图 4.9 所示。首先从双人语音上提取出基于预测差分幅度谱的特征序列（见第二章）；然后按照基于 UBM 的说话人分割聚类算法将双人特征序列分离为两段只含有

单个说话人的特征序列（见第三章）；接着与基于能量的有效音检测的结果进行合并，去除每段特征序列中的非语音部分；最后按照基于信道子空间投影的模型补偿算法（见第四章）对每个说话人的特征序列进行打分，取两者的最大得分作为输出结果。

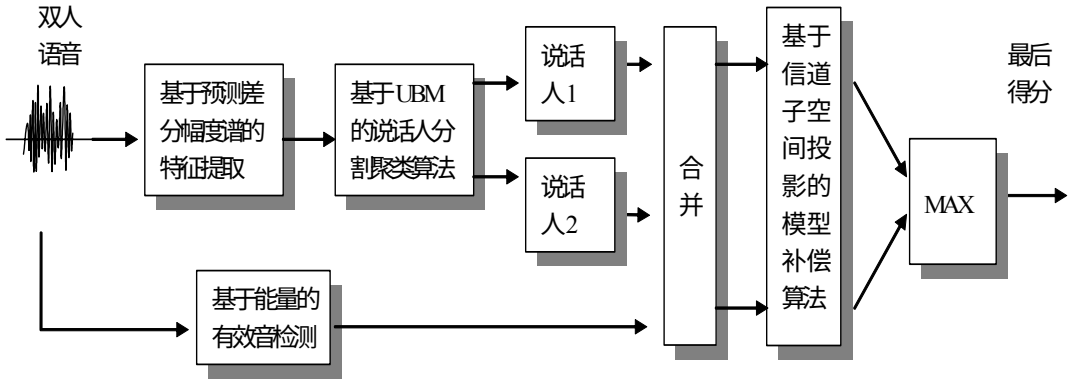

图 4.9 电话信道下双人识别系统示意图

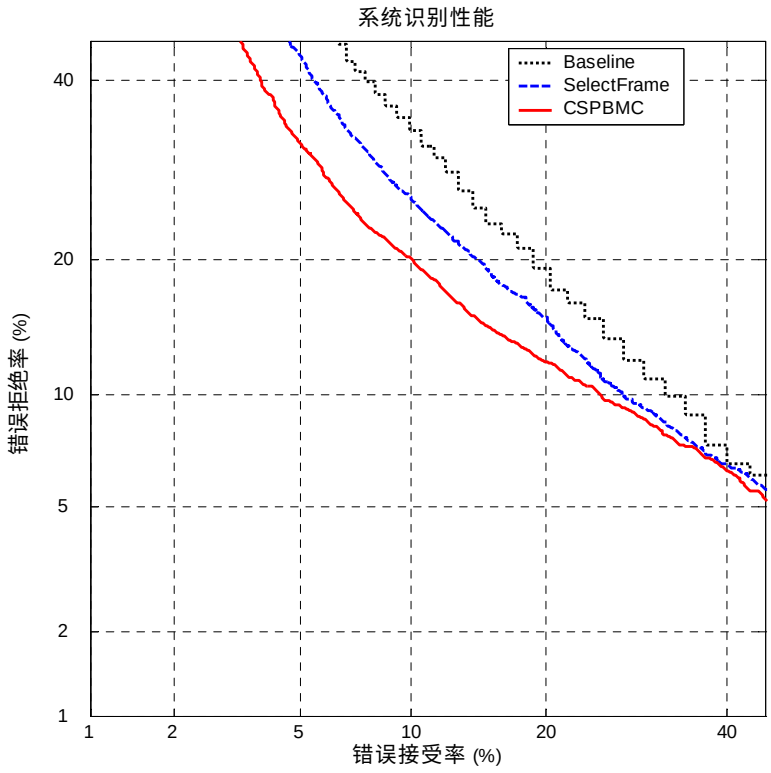

图 4.10 三种系统在双人识别实验中的 DET 曲线对比

实验中对比了三种系统：(1) 在基于 UBM 的说话人分割聚类算法中不带“挑帧处理”，不使用基于信道子空间投影的模型补偿算法，该系统记做 Baseline；(2) 在 Baseline 系统的基础上，基于 UBM 的说话人分割聚类算法里采用“挑帧处理”(见 3.2.4 节)，该系统记做 SelectFrame。这里去除了分数在后 20%的语音帧；(3) 在 SelectFrame 系统的基础上，与基于信道子空间投影的模型补偿算法相结合，该系统记做 CSPBMC。在这三种系统中，均使用了 Tnorm 算法。

表 4.4 三种系统在双人识别实验中的系统性能对比

系统	最小 DCF($\times 10^{-2}$)	EER(%)
Baseline	9.9	19.3
SelectFrame	8.9	17.1
CSPBMC	7.7	14.7

图 4.10 和表 4.4 给出 Baseline、SelectFrame 和 CSPBMC 三种系统在双人识别数据库上的结果对比。从实验结果可以看出，在 SelectFrame 系统中，通过“挑帧处理”，去掉了一些对识别有害的语音帧，提高了系统的识别性能，相对于 Baseline 系统来说，等错误率相对下降了 11.4%；在 CSPBMC 系统中，采用了基于信道子空间投影的模型补偿算法，降低了测试语音与训练语音之间由于信道差异带来的影响，相对于 Baseline 系统来说，等错误率相对下降了 23.8%。从最小 DCF 来看，CSPBMC 系统相对于 Baseline 系统下降了 22.2%。

双人识别系统相对于单人识别系统来说，仍有很大的不足，主要原因有：(1) 多人语音经过分割聚类后，仍然存在有一定的聚类错误，影响了系统的识别性能；(2) 论文里使用了比较简单的挑帧处理方法，在一定程度上提高了系统的识别性能，但仍不能很好的去除语音中对识别有干扰的语音帧；(3) 由于分割聚类错误的存在，使得用于识别的语音段中含有其他信道的语音，这会在一定程度上影响对信道子空间投影的估计。总的来说，单人识别的结果是双人识别的一个上限，它的系统性能会对双人识别系统有较大的影响。在理论上，如果分割聚类算法足够精确，那么双人识别的结果应该跟单人识别相差不多。但由于分割聚类

错误和一些重叠语音的存在，使得双人识别系统的性能差于单人识别系统。从 NIST 2006 年评测来看，双人识别系统的等错误率一般为单人识别的 1.5 倍，说明现有的分割聚类算法在实际应用中仍存在需要改进的地方，如对重叠语音的处理。

4.4 小结

本章针对面向电话信道应用的说话人识别系统面临的信道差异问题，在 LFA 和 NAP 的基础上，提出了一种简单的基于信道子空间投影的模型补偿算法，一方面避免了 LFA 中信道因子的复杂计算，另一方面可以应用于GMM-UBM中，不需要额外的分类器（如NAP中的SVM）。该算法通过补偿的方式，把从测试语音在信道子空间上的投影加到 UBM 和说话人模型上，以便改善测试语音和训练语音之间的信道不匹配。但是由于信道子空间投影是从说话人模型上估计得到的，因此说话人模型的精度对于本文提出的算法有很大的影响的，也就是说测试语音的长度对于本文提出的算法的性能至关重要。从实验结果可以看出，本文的算法在测试语音长度较短的情况下，识别性能有所降低；但是随着测试语音长度的增加，自适应得到的说话人模型精度越来越高，从而能够较准确的估计出测试语音在信道子空间上的投影，使得补偿后的模型与测试语音能够相匹配。另一方面，本章的式（4-19）等号成立的条件是 N 足够的大，但是本章里的 N 为 8，因此会使得式（4-20）中的对 $Uz(s,c_i)$的估计不够准确。如果条件允许的话，使用更大的 N 可能会取得更好的效果。

第5章 总结与展望

5.1 论文工作总结

科学技术的飞速发展使得人们在日常生活中不得不记住各种各样的口令或密码，以便向他人或机器证明自己的合法身份。而这些口令或密码的遗失或忘记，则会给人们带来各种各样的烦恼和损失。因此，生物特征识别便成为一种非常好的解决方案，而说话人识别则是其中的一种。说话人识别的研究始于 20 世纪 30 年代，并在近三十多年里取得了长足的进步。本文针对其中的一个研究任务，即电话信道下多说话人识别中的若干难点和问题，在背景噪音、说话人分割聚类和信道差异方面进行了初步的探索和研究，提出了一些新的方法，并通过实验证明了其有效性，同时也为进一步深入研究打下了一定的基础。

概括来说，本文的工作重点和贡献主要体现在如下几个方面：

1．针对电话语音中存在的背景噪音问题，提出了一种**基于预测差分幅度谱的噪音鲁棒特征提取算法**。该算法利用一帧语音内相邻频率上的幅度差来去除噪音的影响，并与非线性谱减法和基于差分能量谱的噪音鲁棒算法进行了比较分析。由于噪音的影响，会使得干净语音谱中峰谷信息的估计不准确，为了解决这一问题，采用了一个正弦滤波器来预测当前位置所处的峰谷信息。根据得到的峰谷位置，采用不同的加权差分函数来去除背景噪音的影响和保留频谱中的说话人特性，最后通过累积运算来恢复原始干净的语音频谱。这种算法的一个好处是不需要对噪音谱做预先的估计。在四种噪音类型，不同信噪比下，该算法的平均错误率相对于非线性谱减法下降 24.1%。实验结果表明，差分能量谱要比原始能量谱有较好的噪音鲁棒性，而根据峰谷信息的加权差分要好于简单的差分。但是该算法在信噪比较低的情况下，对频谱中峰谷信息的估计错误较多，使得系统的识别率相对于信噪比较高的情况有较大的下降。

2．针对电话交谈语音中短语音段较多的现象，提出了一种**基于**

UBM 的说话人分割聚类算法。该算法可分为三个阶段：初始分割、聚类和重分割。(1) 初始聚类阶段：由于电话交谈语音中每个说话人的语音段有长有短，这就要求分割时使用的“距离”度量准则能够很好的处理短语音段的情况。考虑到在较短时间段内，同一个说话人的两段语音在同一模型上的似然分差异不大，而 UBM 又能够代表大多数说话人的发音特性，因此本文采用了基于 UBM 的对数似然比分来作为说话人分割的一种“距离”度量准则，用来找出多人语音中可能的说话人转换点。为了进一步降低分割错误中的 FAR，便于后面聚类模块的处理，本文使用了 BIC 方法来优化上一步分割的结果，将其中属于同一说话人的相邻两段语音段进行合并。实验中使用了 NIST 2002 Switchboard 数据库，并将提出的算法跟 BIC、GLR 和 DISTBIC 进行了比较分析。该算法的错误率相对于 DISTBIC 来说下降了 13.5%。(2) 聚类阶段：由于电话交谈语音一般只含有两个说话人，因此本文对所研究的问题做了简化，假定多人语音中只含有两个说话人。为了对语音中含有的两个说话人建立正确的说话人模型，并在此基础上对分割后的语音段按照话者的身份进行归类，本文提出了一种基于模型间分数差的聚类方法。该算法将一段语音在两个模型上的分数差作为该语音段属于某个模型的“概率分”，并根据“概率分”的大小，选择得分较大的语音段训练或更新特定说话人的模型。在初始聚类时，由于每段语音的身份未知，并且存在较多的短语音段，因此使用了小混合的 UBM 来生成说话人模型；在初始聚类后，由于可用于训练说话人模型的语音较多，可以使用大混合的 UBM 来得到精确的说话人模型。为了使语音段在两个说话人模型上得分的区分性更加明显，本文使用了 Dnorm 算法，进一步降低了聚类错误率。(3) 重分割阶段：由于聚类是在初始分割的基础上进行了，没有对该结果做任何改动，这会使得在初始分割时产生的漏检错误得不到消除。为此，本文用聚类得到的说话人模型对原始语音进行了重分割处理，使得错误率进一步降低到 4.5%。与 NIST 2002 年电话交谈语音分割聚类评测中性能最好的系统（Fusion+LIA）相比，聚类错误率相对下降了 21.1%。

3．针对电话信道应用中遇到的信道差异问题，提出了一种**基于信道子空间投影的模型补偿算法**，来降低测试语音与训练语音间的信道不匹配。该算法是在超向量空间上，分析信道差异给说话人模型带来的影

响，并将 LFA 中模型补偿的思想与 NAP 中子空间投影的思想结合起来，即用子空间投影的方式得到语音中含有的信道信息，并用该信息对说话人模型进行补偿，来降低测试语音与训练语音之间由信道差异所带来的不匹配。这样，一方面使得对语音中含有的信道信息的估计变得简单，另一方面使得算法能够很好的应用于 GMM-UBM 系统中。该算法包括以下几步：(1) 用 PCA 方法从大量的集外说话人数据上得到超向量空间上，信道子空间的投影矩阵 P；(2) 在训练时，首先用传统的 MAP 算法从 UBM 上得到说话人模型，接着将该模型中每个混合的均值向量连接成一个超向量 m，最后把$(I-P)m$ 作为说话人模型保留下来；(3) 在测试时，首先用传统的 MAP 算法从 UBM 上得到说话人模型，接着将该模型中每个混合的均值向量连接成一个超向量 m'，计算测试语音在信道子空间上的投影向量 Pm'，最后用加上 Pm'后的 UBM 和说话人模型来识别测试语音的话者身份。在 NIST 2006 年单人测试数据库上，与 Tnorm 相结合，系统的等错误率为 9.3%。相对于只用 Tnorm 的系统来说，等错误率相对下降了 16.2%。

5.2 下一步研究的展望

本文虽然在电话信道下多说人识别方面进行了一些初步研究，在前人研究的基础上取得了一些研究成果，但同时也发现了一些不足之处。下面将针对这些不足之处，指出今后计划进一步深入开展研究的若干方向。

1．本文提出的基于预测差分幅度谱的噪音鲁棒算法只是对频谱中峰谷信息和差分幅度谱进行了初步的研究，并且仅仅在几种典型的噪声下测试了算法的性能，离实际应用还有很大的距离。然而在实际生活中，即使受到较强的噪音干扰，人们对语音话者身份的识别能力还是相当鲁棒的，这是因为人在识别话者身份的时候能够综合运用多种信息进行判定，例如习惯用语、语速、口音等等。因此，研究如何提取出含有这些高层信息的特征，如何将提取出的高层信息相关的特征与现有低层声学特征想结合，以及如何将语音识别与说话人识别两种技术结合起来将是提高说话人识别系统鲁棒性的途径之一，也是今后的一个研究方向和内

容。

2．本文研究的说话人分割聚类算法仅仅是在前人研究的基础上做了初步的探索，并对研究的问题做了简化处理，假定多人语音中仅含有两个说话人。这与实际应用中的情况是不符合的，因此如何判断出一段多人语音中含有的话者数目是今后的一个研究内容。另外，本文对于重叠语音的处理是非常简单的，并不能保证将语音中的重叠语音段完全挑出来，这样就会影响后面的识别模块，使得系统不能达到单人识别系统所能达到的识别性能。因此，如何更好的将重叠语音段挑选出来是今后的一个研究内容。

3．由于先验信道数据的缺乏，使得本文的系统无法进行特征映射处理。如果在条件允许的情况下，能够进行特征映射，那么（1）在进行说话人分割聚类时，可以按照分窗的方式，先对每窗内的语音进行信道类型判别，然后进行特征映射处理，这样就能够在一定程度上消除多人语音中的信道差异，有可能会进一步提高分割聚类算法的性能；(2) 在进行分数归一化时，能够进一步降低系统的识别错误率。那么在得到大量先验信道数据的前提下，如何保证训练得到的信道相关 UBM 对信道分类的正确率呢？作者在用已知信道类型的数据训练信道相关的 UBM 时发现，即使在同一信道下，特征的分布仍然比较松散，这就会使 UBM 不具备好的信道分类性能。可能的一个解决方案是对分类性能不好的 UBM，将其训练数据按一定的聚类算法进一步分为许多小类，以便得到多个 UBM，然，后用这些 UBM 来进行信道分类，或许能够提高信道分类的准确率，这也将是今后的一个研究内容。

参考文献

[1] Pruzansky S. Pattern-matching procedure for automatic talker recognition. Journal of the Acoustical Society of America. 1963, 35(3):354-358

[2] Atal B S. Automatic recognition of speakers from their voices. Proc. IEEE. 1976, 64(4):460-475

[3] Davis S B and Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustic, Speech and Signal Processing. 1980, 28:357-366

[4] 甄斌，吴玺宏，刘志敏，迟惠生. 语音识别和说话人识别中各倒谱分量的相对重要性. 北京大学学报（自然科学版），2001，37(3):371-378

[5] Hermansky H. Perceptual linear prediction (PLP) analysis for speech. Journal of the Acoustic Society of America (JASA). 1990, 87(4):1738-1752

[6] 郭庆. 声学模型中帧间相关性和自适应问题的研究. [博士学位论文]. 北京, 清华大学计算机系, 1999

[7] Furui S. Comparison of speaker recognition methods using static features and dynamic features. IEEE Transaction on Acoustics, Speech, and Signal Processing. June 1981. 29(3):342-350

[8] 段新，黄新宇，吴淑珍. 与文本无关的说话人辨认系统中一种新的使用基音周期方法研究. 北京大学学报，2003，39(5): 690-696

[9] 林玮，杨莉莉，徐柏龄. 基于修正 MFCC 参数汉语耳语语音的话者识别. 南京大学学报，2006，42(1): 54-62

[10] Reynolds D A, Campbell W, Gleason T T, et al. The 2004 MIT Lincoln laboratory speaker recognition system. In Proceedings of ICASSP. Philadelphia, USA, 2005, 177-180

[11] Chen Z H, Liao Y F and Juang Y T. Prosody modeling and eigen-prosody analysis for robust speaker recognition. in Proceedings of ICASSP. Philadelphia, USA, 2005, 185-188

[12] Adami A G. Prosodic modeling for speaker recognition based on sub-band energy temporal trajectories. in Proceedings of ICASSP. Philadelphia, USA, 2005, 189-192

[13] Mijail A, Anil A, Philip Z, et al. A Bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition. in Proceedings of InterSpeech. Lisbon, Portugal,

2005. 2009-2013

[14] Ramachandran R P, Farrell K R, Ramachandran R, et al. Speaker recognition-general classifier approaches and data fusion methods. Pattern Recognition. December 2002, 35(12):2801-2821

[15] 杨行峻，迟惠生，等. 《语音信号数字处理》. 电子工业出版社，1995 年

[16] Hertz J, Krogh A and Palmer R G. Introduction to the theory of neural computation. Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, Reading, Mass, USA. 1991

[17] Haykin S. Neural networks: a comprehensive foundation. Macmillan, New York, NY, USA, 1994

[18] Vapnik V N. The nature of statistical learning theory. Springer-Verlag, New York, 1995

[19] Sakoe H and Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustic, Speech and Signal Processing. 1978, ASSP-26(1):43-49

[20] Higgins A L, Bahler L G and Porter J E. Voice identification using nearest neighbor distance measure. in Proceedings of ICASSP. 1993. 375-378

[21] Rabiner L R and Juang B H. Fundamentals of speech recognition. Signal Processing. Prentice-Hall, NJ, 1993

[22] 刘鸣，戴蓓倩，李辉，等. 鲁棒性话者辨识中的一种改进的马尔可夫模型. 电子学报，2002，30(1):46-48

[23] 朱晓园. 一个对隐马尔可夫模型用于自由语句说话人的研究. 北方交通大学学报，1997，21(1):34-38

[24] Reynolds D A and Rose R C. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing. January 1995, 3(1):72-83

[25] 马继涌，高文. 基于最大交叉熵估计高斯混合模型参数的方法. 软件学报，1999，10(9):974-978

[26] Gish H and Schmidt M. Text-independent speaker identification. IEEE Signal Processing Mag. 1994, 11:18-32

[27] Reynolds D A. Comparison of background normalization methods for text-independent speaker verification. in Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997). Rhodes, Greece, 1997, 2: 963-966

[28] Reynolds D A, Quatieri T F and Dunn R B. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing. 2000, 10(1-3):19-41

[29] Bennani Y and Gallinari P. On the use of TDNN-extracted features information in talker identification. International Conference on Acoustics, Speech and Signal Processing. 1991, 385-388

[30] Farrell K P, Mammone R J and Assaleh K T. Speaker recognition using neural networks and conventional classiers. IEEE transactions on Speech and Audio Processing. 1994, (2):194-205

[31] Schmidt M and Gish H. Speaker identification via support vector classifiers. in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'96), Atlanta, Ga, USA, May 1996, (1): 105-108

[32] Gu Y and Thomas T. A text-independent speaker verification system using support vector machines classifier. in Proc. European Conference on Speech Communication and Technology (Eurospeech 2001). Aalborg, Denmark, September 2001, 1765-1769

[33] Liu M H, Dai B Q, Xie Y L and Yao Zh Q. Improved GMM-UBM/SVM for speaker verification. ICASSP 2006, 1:925-928

[34] Kharroubi J, Petrovska D D and Chollet G. Combining GMMs with support vector machines for textindependent speaker verification. in Proc. European Conference on Speech Communication and Technology (Eurospeech 2001). Aalborg, Denmark, September 2001, 1757-1760

[35] Xin D, Wu Zh H and Yang Y Ch. Exploiting support vector machines in hidden Markov models for speaker verification. in Proc. 7th International Conf. on Spoken Language Processing (ICSLP 2002), Denver, Colo, USA, September 2002, 1329-1332

[36] Lee C H, Lin C H and Juang B H. A study on speaker adaptation of parameters of continuous density hidden Markov models. IEEE Trans. on Acoustic and Speech Signal Processing. 1991, 39(4): 806-814

[37] Lee C H and Gauvain J L. Speaker adaptation based on MAP estimation of HMM parameters. in Proc. Int. Conf. Acoustics, Speech, and Signal Processing. 1993, 2: 652-655

[38] 李虎生，刘加，刘润生. 语音识别说话人自适应研究现状及发展趋势. 电子学报，2003，31(1): 103-108

[39] Leggetter C J and Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language. 1995, 9:171-185

[40] Leggetter C J and Woodland P C. Flexible speaker adaptation for large vocabulary speech recognition. in Proc. of Eurospeech 1995, 1155-1158

[41] Leggetter C J. Improved acoustic modeling for HMMs using linear transformations. PhD thesis. Cambridge University. 1995

[42] Chen S S and Gopalakrishnan P S. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. in DARPA Speech Recognition Workshop, 1998

[43] Moraru D, Meignier S, Besacier L, et al. The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation. in Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, 2003, 2: 89-92

[44] Moraru D, Meignier S, Fredouille C, et al. The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. in Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2004), Montreal, Canada, 2004, 373-376

[45] Fredouille C, Moraru D, Meignier S, et al. The NIST 2004 spring rich transcription evaluation: two-axis merging strategy in the context of multiple distance microphone based meeting speaker segmentation. in RT2004 Spring Meeting Recognition Workshop, 2004, 5-8

[46] Ivan M C, Aaron E R and Parthasarathy S. Detection of target speakers in audio databases. in Proceedings of ICASSP 1999, Phoenix, Arizona, United States. 821-824

[47] Wu T, Lu L, Chen K, Zhang H J. UBM-based real-time speaker segmentation for broadcasting news. in Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP 2003), Hong Kong, China, 2003, (2):193-196

[48] Delacourt P and Wellekens C J. DISTBIC: a speaker-based segmentation for audio data indexing. Speech Communication, Sept. 2000, 32(1-2):111-126

[49] Gish H, Siu M H and Rohlicek R. Segregation of speakers for speech recognition and speaker identification. in IEEE International Conference on Acoustics Speech and Signal Processing. 1991, 873-876

[50] Siegler M A, Jain U, Raj B and Stern R M. Automatic segmentation classification and clustering of broadcast news audio. in DARPA Speech Recognition Workshop, 1997, 97-99

[51] Yi G X, Li Q Y, Lin Zh S, Wu X H, Chi H Sh. Speaker segmentation based on model scoring. Technical Acoustics 2005, (24):218-221

[52] 白俊梅，张树武，徐波. 广播电视中的目标说话人跟踪技术. 声学技术 2005，(24):234-238

[53] 吕萍，颜永红. 广播新闻语料自动识别系统. 声学技术，2005，(24):109-112

[54] Boll S F. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing. 1979, 27:113-120

[55] Berouti M, Schwartz R and Makhoul J. Enhancement of speech corrupted by acoustic noise. ICASSP, 1979, 208-211

[56] 田斌，易克初. 一种用于强噪声环境下语音识别的含噪 Lombard 及 Loud 语音补偿方法. 声学学报，2003，28(1): 28-32

[57] Lockwood P and Boudy J. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection of robust speech recognition in cars. Speech Communication, 1992. 11(6): 215-228

[58] Poruba J. Speech enhancement based on nonlinear spectral subtraction. in IEEE Proc. Int. Conf. on Device, Circuits and Systems, 2002, 546-549

[59] Hermansky H and Morgan N. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, 1994. 2(4): 578-589

[60] 吕成国，王承发，李俊庆，等. RASTA-PLP 技术与谱减法相结合的去噪方法. 自动化学报，2000，26(5): 717-720

[61] Zhen B, Wu X H, Liu Z M, Chi H S. An enhanced RASTA processing for speech signal. Chinese Journal of Acoustics, 2001, 26(3): 252-258.

[62] 陈景车，姚磊，黄泰翼. 几种高鲁棒性通信及说话人自适应语音识别算法研究. 声学学报，1998，23 (6) :573-544

[63] Kocsor A, Toth L, Kuba A, et al. A comparative study of several feature transformation and learning methods for phoneme classification. International Journal of Speech Technology, 2000. 3(3): 263-276

[64] Saon G, Padmanabhan M, Gopinath R, Chen S. Maximum likelihood discriminant feature spaces. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2000. 2: 1129-1132

[65] Gales M J F and Young S J. Robust continuous speech recognition using parallel model combination. IEEE Transactions on Speech and Audio Processing, 1996, 4(5): 352-359

[66] Renevey P and Drygajlo A. Statistical estimation of unreliable features for robust speech recognition. in Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000), Istambul, Turkey, 2000, 1731-1734

[67] Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Processing, 1981. 29(2):254-272

[68] Pelecanos J and Sridharan S. Feature warping for robust speaker verification. in Proc. Speaker Odyssey 2001 conference, June 2001, pp.213-218

[69] Reynolds D A. Channel robust speaker verification via feature mapping. in

ICASSP, 2003, (2): 53-56

[70] Teunen R, Shahshahani B and Heck L P. A modelbased transformational approach to robust speaker recognition. In Proc. in ICSLP, 2000, 213-218

[71] 周静芳，陈一宁，刘加，刘润生. 说话人识别信道补偿技术 HNSSM. 清华大学学报（自然科学版）2004，24(7):942-945

[72] Vogt R and Sridharan S. Experiments in session variability modeling for speaker verification. ICASSP, Toulouse, France, May 2006. 897-900

[73] Kenny P, Boulianne G and Dumouchel P. Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing, 2005. 13(3):345-354

[74] Solomonoff A, Campbell W and Boardman I. Advances in channel compensation for SVM speaker recognition. ICASSP, 2005. 629-632

[75] Campbell W M, Sturim D E, Reynolds D A and Solomonoff A. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. ICASSP, 2006. 97-100

[76] Reynolds D A. The effect of handset variability on speaker recognition performance: experiments on the switchboard corpus. in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP 1996), Atlanta, Ga, USA, May 1996. 113-116

[77] 陈皓，付中华，赵荣椿. 说话人确认中针对语音编码差异的似然比得分补偿方法. 西北工业大学学报 2005，23(4):534-537

[78] Lu L, Zhang H J and Jiang H. Content analysis for audio classification and segmentation. IEEE Trans. on Speech and Audio Processing, Oct. 2002. 7(10): 504-516

[79] Naik J M, Netsch L P and Doddington G R. Speaker verification over long distance telephone lines. ICASSP 1989. 524-527

[80] Higgins A, Bahler L and Porter J. Speaker verification using randomized phrase prompting. Digital Signal Processing 1991, 1(2):89-106

[81] 秦兵，陈惠鹏，李光琪，刘松波. 文本有关的话者确认系统. 哈尔滨工业大学学报，2000，32(4):16-18

[82] 邓浩江，杜利民，万洪杰. 似然得分归一化及其在文本无关说话人确认中的应用. 电子与信息学报，2005，27(7):1025-1029

[83] Li K P and Porter J E. Normalizations and selection of speech segments for speaker recognition scoring. in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP 1988), New York, NY, USA, April 1988. 595-598

[84] Auckenthaler R, Carey M and Lloyd-Thomas H. Score normalization for

text-independent speaker verification system. Digital Signal Processing, 2000. 10(1-3):42-54

[85] Ben M, Blouet R and Bimbot F. A Monte-Carlomethod for score normalization in automatic speaker verification using Kullback-Leibler distances. in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP 2002), Orlando, Fla, USA, May 2002. 689-692

[86] Adami A, Mihaescu R, Reynolds D and Godfrey J. Modeling prosodic dynamics for speaker recognition. ICASSP, 2003. 106-110

[87] Peskin B, Navratil J, Abramson J, et al. Using prosodic and conversational features for high-performance speaker recognition. Report from JHU WS'02, ICASSP, 2003, 792-795

[88] Andrews W D, Kohler M A, Campbell J P, et al. Gender-dependent phonetic refraction for speaker recognition. ICASSP, 2002, 149-152

[89] Klusacek D, Navratil J, Reynolds D A, et al. Conditional pronunciation modeling in speaker detection. ICASSP, 2003, 804-807

[90] Doddington G. Speaker recognition based on idiolectal differences between speakers. Eurospeech, 2001. 4:2521-2524

[91] Chen J D, Paliwal K K and Nakamura S. Cepstrum derived from differentiated power spectrum for robust speech recognition. Speech Communication, 2003. 41:469-484

[92] 田滨，曹志刚. 帧间约束 MMSE 语音增强算法. 电子学报 1995，23(9):12-18

[93] 黄磊，吴顺君，张林让，冯大政. 快速子空间分解方法及其维数的快速估计. 电子学报，2005，33(6): 977-981

[94] 罗宇，杜利民. 基于概率加权平均的 Mel 子带特征重建算法. 电子学报，2004，32(10): 1738-1741

[95] Viikki O and Laurila K. Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization. in ESCA NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, 1997, 107-110

[96] Moon S Y and Hwang J N. Noisy speech recognition via wavelet coefficient enhancement. in Proc. IEEE 26th Asilomar Conf. Signals, Syst., Comput., Monterey, CA, Oct. 1992, 1086-1090

[97] 陶智，赵鹤鸣和龚呈卉. 基于听觉掩蔽效应和 Bark 子波变换的语音增强. 声学学报（中文）2005，30(4):367-372

[98] 吴礼福，姚志强，戴蓓蒨和李辉. 音源特征用于提高话者确认系统的鲁棒性. 中国科学技术大学学报 2005，36(5):476-480

[99] Carlson B A and Clements M A. A projection-based likelihood measure for speech recognition in noise. IEEE Transactions on Speech and Audio Processing 1994, 2:97-102

[100] Strope B and Alwan A. A model of dynamic auditory perception and its application to robust word recognition. IEEE Transactions on Speech and Audio Processing 5, 1997, 451-464

[101] Varga A P, Steeneken H J M, Tomlinson M and Jones D. The noisex-92 study on the effect of additive noise on automatic speech recognition. Technical report, Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK., 1992

[102] Rissanen J. Stochastic complexity in statistical inquiry. Series in Computer Science, 1989, Vol. 15. World Scientific, Singapore, Chapter 3

[103] Sivakumaran P, Fortuna J and Ariyaeeinia A M. On the use of the Bayesian information criterion in multiple speaker detection. In Eurospeech, 2001, 795-798

[104] Sivakumaran P, Ariyaeeinia A M and Fortuna J. An effective unsupervised scheme for multiple-speaker-change detection. ICSLP 2002, 569-572

[105] Tritschler A and Gopinath R. Improved speaker segmentation and segments clustering using the Bayesian information criterion, Proc. of Eurospeech, 1999. 679-682

[106] Campbell J P. Speaker recognition: a tutorial. Proc. IEEE, 1997. 9(85):1437-1462

[107] Bonastre J F, Delacourt P, Fredouille C, et al. A speaker tracking system based on speaker turn detection for NIST evaluation. in Proc of ICASSP 2000, Istanbul, Turkey. 2000. 1177-1180

[108] Liu D and Kubala F. Fast speaker change detection for broadcast news transcription and indexing. in Proc. of Eurospeech, 1999, 1031-1034

[109] Bimbot F, Magrin-Chagnolleau I and Mathan L. Secondorder statistical measures for text-independent speaker identification. Speech Communication, 1995. 17(1-2): 177-192

[110] Reynolds D A, Singer E, Carlson B A, McLaughlin J J, et al. Blind clustering of speech utterances based on speaker and language characteristics. in Proceedings of ICSLP 1998. 610-613

[111] Meignier S, Bonastre J F and Magrin-Chagnolleau I. Speaker utterances tying among speaker segmented audio documents using hierarchical classification: towards speaker indexing of audio databases. in Proceedings of ICSLP 2002, Denver, Colorado, United States, September 2002. 1:573-576

[112] Meignier S, Bonastre J F and Igounet S. E-HMM approach for learning and

adapting sound models for speaker indexing. in 2001: A Speaker Odyssey, Chania, Crete, June 2001. 175-180

[113] http://www.nist.gov/speech/tests/spk/2002/resource/index.htm

[114] Xiong Zh Y, Zheng T F, Song Zh J and Wu W H. Combining selection tree with observation reordering pruning for efficient speaker identification using GMM-UBM. Proc. ICASSP. 2005, 625-628

[115] Mathieu B, Frédéric B and Guillaume G. Enhancing the robustness of Bayesian methods for text-independent automatic speaker verification. in ODYS-2004, 165-172

[116] Barras C and Gauvain J L. Feature and score normalization for speaker verification of cellular data, in proc of ICASSP 2003, 2:49-52

[117] Kenny P, Boulianne G, Ouellet P and Dumouchel P. Factor analysis simplified. ICASSP 2005, 637-640

[118] Campbell B, Sturim D E, Shen W, et al. MIT Lincoln laboratory site presentation. NIST speaker recognition workshop 2006

[119] Sturim D E and Reynolds D A. Speaker adaptive cohort selection for Tnorm in text-independent speaker verification. ICASSp 2005, pp.741-744

[120] Martin A, Doddington G, Kamm T, et al. The DET curve in assessment of detection task performance. in Proc. European Conference on Speech Communication and Technology (Eurospeech 1997), Rhodes, Greece, September 1997. (4): 1895 – 1898

[121] http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf

致　谢

衷心感谢我的导师吴文虎教授和郑方教授四年来对我的悉心指导和关怀。两位导师严谨求实，平易近人，无论在学业上还是在生活上都给予了我莫大的关心和帮助。在此，谨向两位恩师致以最诚挚的谢意！

感谢语音技术中心的其它老师，包括方棣棠教授、李树青教授、徐明星老师、邬晓钧老师，以及实验室的全体同窗，他们与我进行了许多有益的讨论，同时给予我很多工作上的支持，在此一并向他们表示感谢。

最后，衷心感谢我的家人和朋友，他们无私的爱和默默的关怀，一直伴随着我的奋斗过程。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1978 年 5 月出生于山东省淄博市。

1995 年 9 月考入山东科技大学计算机应用与技术系，1999 年 7 月本科毕业并获得工学学士学位。

1999 年 9 月免试进入山东科技大学计算机应用与技术系攻读硕士，2002 年 7 月硕士毕业获得工学硕士学位。

2002 年 9 月考入清华大学计算机系攻读博士至今。

发表的学术论文

[1] Deng J, Zheng F, Liu J, et al. The predictive differential amplitude spectrum for robust speaker recognition in stationary noises, in Proceeding of the 9th European Conference on Speech Communication and Technology (Interspeech). Lisbon Portugal 2005. 3105-3108

[2] Deng J, Zheng F, Song Zh J, et al. Modeling high-level information by using gaussian mixture correlation for GMM-UBM based speaker recognition, in Proceeding of the 9th European Conference on Speech Communication and Technology (Interspeech). Lisbon Portugal 2005. 2033-2036

[3] Deng J, Zheng F, Song Zh J, et al. Using predictive differential power spectrum and subband Mel-spectrum centroid for robust speaker recognition in stationary noises, in Proceeding of the 4th International Conference on Machine Learning and Cybernetics (ICMLC), Guangzhou China 2005. 4846-4851 (EI indexed, accession number: 05509539585)

[4] 邓菁，郑方，刘建，吴文虎. Mel 子带谱质心和高斯混合相关性在鲁棒话者识别中的应用. 声学学报. 2006. 31(9):471-475（EI 检索源）

被录用的学术论文

[1] Deng J, Zheng F and Wu W H, UBM based speaker segmentation and clustering for 2-speaker detection, ISCSLP 2006, LNAI 4274, 116-125 (To be SCI indexed)

其他学术论文

[1] Liu J, Zheng F, Deng J, et al. Real-time pitch tracking based on combined SMDSF, in Proceeding of the 9th European Conference on Speech Communication and Technology (Interspeech). Lisbon Portugal 2005. 301-304

[2] Chen D F, Zheng F, Liu J, Deng J, et al. The dynamically-adjustable histogram pruning method for embedded voice dialing, in Proceeding of the 7th IASTED International Conference on Signal and Image Processing (SIP). Hawaii, USA 2005. 46-51

在读期间完成的其它研发工作

1. 设计开发声纹识别 API v3.0，并投入实际应用，通过了公安部鉴定（3 人合作完成）
2. 完成可用于多机并行的说话人识别软件，并投入实际应用（2 人合作）
3. 设计开发了 Nokia 6600 手机上的得意整句输入法（2 人合作）
4. 设计开发了用于 PDA 上的文本相关的说话人识别系统（2 人合作）